CatSD: Structural Database and High-Throughput Predictive Workflows for Homogeneous Catalyst Design

Marc Andrew Stephen Short

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds

Faculty of Engineering and Physical Sciences

School of Chemical Engineering

November 2022

Intellectual Property

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement. The right of Marc Short to be identified as the author of this work has been asserted by Marc Short in accordance with the Copyright, Designs and Patents Act 1988.

Signed

Acknowledgements

I would like to thank my supervisor Dr Bao Nguyen for his constant advice, guidance and insight as well as his extensive chemical knowledge which was invaluable for the project.

I would also like to thank my other supervisors Dr Charlotte Willans and Dr Clare Tovee (CCDC) for their support, insight and guidance during the course of the project. Furthermore, I would also like to thank Dr Samuel Boobier for his assistance with the machine-learning models.

I would also like to acknowledge the contribution of our Master's student Zeshen Wang for performing the synthesis of the ligand candidates described in this thesis as well as Dr Tom Nicholls from the Willans groups for performing the experimental validation work during the COVID-19 lockdown.

I would also like to acknowledge that this work was undertaken on ARC3 and ARC4, part of the High-Performance Computing facilities at the University of Leeds, UK. As well as the support of the ARC team for their continued assistance with software and code optimisation during the course of the project.

Thank you to the Centre for Doctoral Training in Complex Particulate Products and Processes (cP^3), the Engineering and Physical Sciences Research Council (EPSRC) and the Cambridge Crystallographic Data Centre for providing the funding for the project.

Abstract

Identification of highly active catalysts is an important process across multiple industries including drug development, process chemistry and agrochemicals. The lack of understanding of ligand properties and catalytic pathways are limiting factors for the uptake of more sustainable and highly active catalysts. Herein we report a novel method for the identification of ligands and the prediction of their activity for homogeneous catalysts from the Cambridge Structural Database. We present CatSD, a structural database complete with catalytically relevant features to enable the mining of organometallic ligands from the CSD. We also present a high-throughput computational workflow for the prediction of activation energies and mechanistic exploration. This workflow is on a timescale similar to experimental high-throughput screening and provides energies with an accuracy of 3.9 $\mathrm{kcal\,mol^{-1}}$. CatSD and the prediction workflow were applied to the Ullmann-Goldberg reaction to identify novel ligands for amine and amide coupling partners. Over 10,000 ligands were identified from the CSD for both coupling partners. The workflow showed excellent reliability for the generation of starting structures (99.7%) and good reliability for the optimisation of important intermediates (>84%) and transition states (TSOA: 33-61%, TSSig: 83-85%). Several ligands were validated experimentally identifying a previously unreported active ligand class. The effect of ligand properties was explored using machine learning to identify several key characteristics for both nucleophile coupling partners. Machine learning was also used to predict activation energies without the need to calculate the transition state. Models were optimised providing accuracy on par with the accuracy of the workflow calculations. It is our hope that the methodologies presented in this work will aid the discovery and design of ligands for homogeneous catalysts for the wider chemistry community as well as stimulate further research in this field.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
AO	Atomic Orbital
Bag	Bagging
CREST	ConformerRotamer Ensemble Sampling Tool
CSD	Cambridge Structural Database
DCM	Dichloromethane
DFT	Density Functional Theory
DLPNO-CCSD(T)	Domain Based Local Pair Natural Orbitals - Coupled Cluster with Single Double and Pertubative Triple Excitations
DMF	N,N-dimethylformamide
DMSO	Dimethylsulfoxide
E_A	Activation Energy
ECP	Effective Core Potential
ET	ExtraTrees
gbsa	Generalised Born model with surface area contributions
GGA	Generalised Gradient Approximation
GP	Gaussian Process
HF	Hartree-Fock
IRC	Intrinsic Reaction Coordinate

LDA	Local Density Approximation
MAE	Mean Absolute Error
ML	Machine Learning
MO	Molecular Orbital
MW	Molecular Weight
PCA	Principal Component Analysis
PLS	Partial Least Squares
Q2MM	Quantum to Molecular Mechanics
QM	Quantum Mechanical
RF	Random Forest
rmsd	Root Mean Squared Deviation
RMSE	Root Mean Square Error
SASA	Solvent Accessible Surface Area
SCF	Self-consistent Field
SMARTS	SMILES arbitrary target specification
SMD	Solvent Model Based on Density
SMILES	Simplified Molecular Input Line Entry System
SQM	Semi-empirical Quantum Mechanical
SVM	Support Vector Machine
TSFF	Transition State Force Field
xTB	Extended Tight Binding

Chapter 1: Introduction

Organometallic catalysis has emerged as a powerful tool in organic chemistry. The role of a catalyst is to provide an alternative reaction pathway that has a lower activation energy (E_A) and therefore increased reaction rate, whilst not being consumed in the process.[1] Catalysts can be found in nature in the form of enzymes, in academic laboratories for research and in industries such as petrochemicals, renewables and pharmaceuticals. Approximately 90% of all chemical products include at least one step using a homo- or heterogeneous catalyst in the manufacturing process.[2]

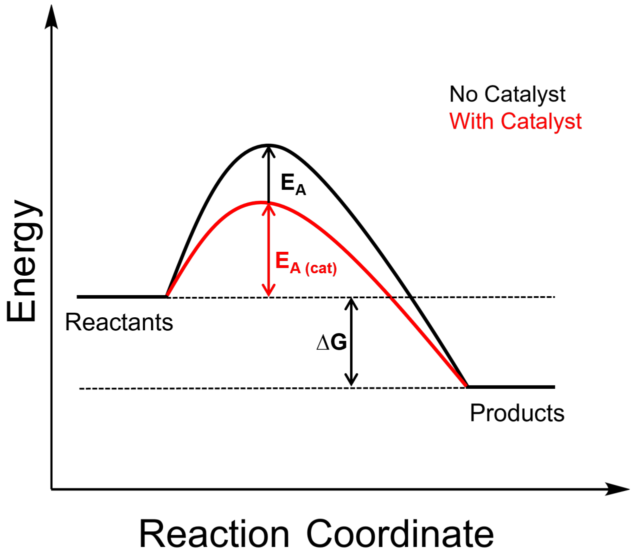


Figure 1.1: Energy profile of a chemical reaction both with (red) and without (black) a catalyst.

Although significant improvements have been made towards increasing substrate scope, functional group tolerance, lower catalyst loadings and reaction conditions for a large variety of synthetically relevant chemical transformations, there is still progress to be made in terms of improving atom economy, cost and reducing the use of toxic compounds during chemical synthesis as well as making more economically viable transition metals available for the wider synthetic toolbox. This could be achieved either through the improvement of existing catalysts or the development of novel catalytic systems.

1.1 Computational Design in Chemistry

A core interest in chemistry is the creation of specific chemical structures and even more importantly chemical structures with specific functions. Understanding the connection between chemical structure and activity allows chemists to make rational decisions to modify and improve performance. Structure-Activity Relationships (SAR) are valued across a breadth of chemical disciplines ranging from catalysis and materials science to synthesis and biology. Such an approach underpins physical organic chemistry whereby chemists aim to characterize the molecular structure and rationalize activity via experimental and theoretical approaches. Theoretical tools, namely computational chemistry, allow the study of molecular structures without the need for physical matter. Computational chemistry enables the investigation into chemical behaviour via simulation, the quantification of which allows for the prediction of desired properties and reactivity as well as the potential for the guided design of highly active compounds.

Computational tools available today are based on a range of physical models of varying degrees of sophistication and accuracy (**Figure 1.2**). Methods such as molecular mechanics and semi-empirical methods are computationally cheaper and allow the mapping of the chemical landscape in an approximate sense. More advanced methods such as density functional theory, Møller-Plesset perturbation theory and coupled cluster allow much more accurate mapping of the energy surface granting access to more 'advanced' properties such as where bonds are made and broken, electronic properties, excited states, hyperfine coupling constants and zero-field splitting.[3]

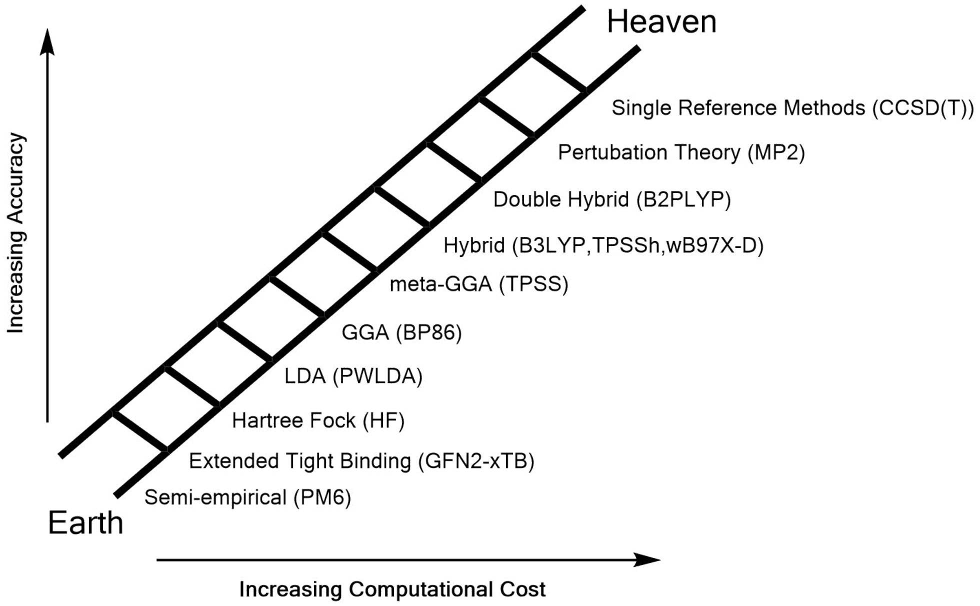


Figure 1.2: Ladder of available computational methods with increasing chemical accuracy.

Along with the rapid advancement of theoretical methods and their computational implementations, both hardware and software capabilities have also advanced at a similar pace. Computational assessment of many properties can now be done on a time scale comparable to experiment. For example, in 1980 the transition state calculation for the Diels-Alder cycloaddition between butadiene and ethylene required approximately 6 months of computational time.[4] Today the same calculation at the Hartree-Fock level of theory takes under a minute.[4] Given these advancements the amount of new catalysts emerging from computational design is still extremely low.[5]

1.1.1 Why Should we Design Catalysts Computationally?

Catalysis has become an integral component of modern-day chemistry from chemical synthesis to catalytic converters in cars. Homogeneous catalysis particularly has transformed the way synthetic chemists make target molecules with the development of advanced transition-metal-based catalysts making new chemical transformations possible.

Chemists are often presented with surprising and unpredictable reactivity. The ability to understand these unpredictable cases is key to being able to manipulate them to develop solutions to known problems or to unlock new applications. Computational methods are one way of trying to understand these complex problems, however, this is no easy task.

Catalytic mechanisms are often very complex, proceeding through short-lived often unstable intermediates at low concentrations, making the isolation and/or characterization of these species experimentally extremely difficult and sometimes impossible. While the isolation of such intermediates is sometimes possible, often they are byproducts representing, deactivation pathways, catalytically inactive species or prevalent side reactions. During catalytic pathways the active metal centre may change between different oxidation states, charges, coordinating geometries and spin states. In most cases the metal centre will also carry multiple ligands to steer the reactivity and selectivity of the reaction, introducing more potential pathways involving ligand exchange which can lead to side reactions or deactivation of the catalyst.

The very nature of catalysts makes designing new catalysts from scratch a challenging task. As catalysts aim to reduce the energetic cost of a reaction, this leads to a flatter potential energy surface compared to the uncatalysed process. A flatter potential energy surface is more likely to be perturbed by factors such as additives or solvents, leading to undesired effects such as catalyst degradation, increased competition of side reactions and a change in stereo- or regio-selectivity.[5] The neglect of unanticipated interactions between reaction components can lead to false assumptions, leading to a waste of both time and resources for a catalyst that is doomed to fail, or could potentially lead to the discarding of promising candidates.

Designing catalysts can therefore be considered as either designing a catalyst to fit within an existing system or designing a new system from scratch, the latter being extremely complex. A series of feasible mechanisms need to be identified that cover both productive and unproductive pathways over a large range of energetically accessible conformations, consisting of different oxidation states, spin states and ligation states. This is clearly an enormous task, which is guided by the current understanding of chemical behaviour and the availability of suitable theoretical models. Given current computational power and still rather limited understanding of a wide range of commonly used synthetic transformations, such an approach could be considered unfeasible. It is more manageable to predict relative reactivity or selectivity from within a confined structural domain.

When evaluating potential modes of reactivity, first ground state and transition state geometries must be found. The success of which is highly dependent upon the input given by the chemist. A lack of mechanistic or structural understanding can lead to incorrect inputs, ultimately leading to the failure of the calculations. With each set of calculations an individual reaction step is assessed, the investigation of which will only be as exhaustive as our ability to ask the correct questions. Pathways which have not previously been thought of or intentionally looked for will be much harder to find or missed entirely. Although new reaction path methods such as Nudged Elastic Band and Growing String are able to potentially find new low energy paths between reactants and products.[6,7] Transition state theory alone may also not explain reactivity and selectivity. Dynamic effects such as solvent and counter-ion reorganisation, as well as deactivation pathways, may also play an important role.[8]

The accuracy of available computational models must also be considered. While methods exist to accurately calculate energies of chemical structures, namely coupled-cluster methods, these 'gold standard' methods are extremely computationally expensive in both time, CPU cores and memory usage. The large number of atoms and therefore, electrons present in transition metal complexes make coupled-cluster methods far too expensive for exploring whole catalytic cycles, even with new approximation methods such as the domain-based local pair natural orbital approximation, which reduces the cost down to that of a similar density functional theory and scales linearly with system size.[9] Such methods cannot be realistically applied to a large number of possible structures without extensive computational infrastructure. No single computational method is universally applicable and benchmarking less expensive methods against higher-level methods is a time-consuming, but often necessary process. Relatively small energy differences ($\sim$2 kcal mol^{-1}) can dramatically alter the selectivity of a reaction, therefore, choosing a method that performs well for the system of interest is extremely important.

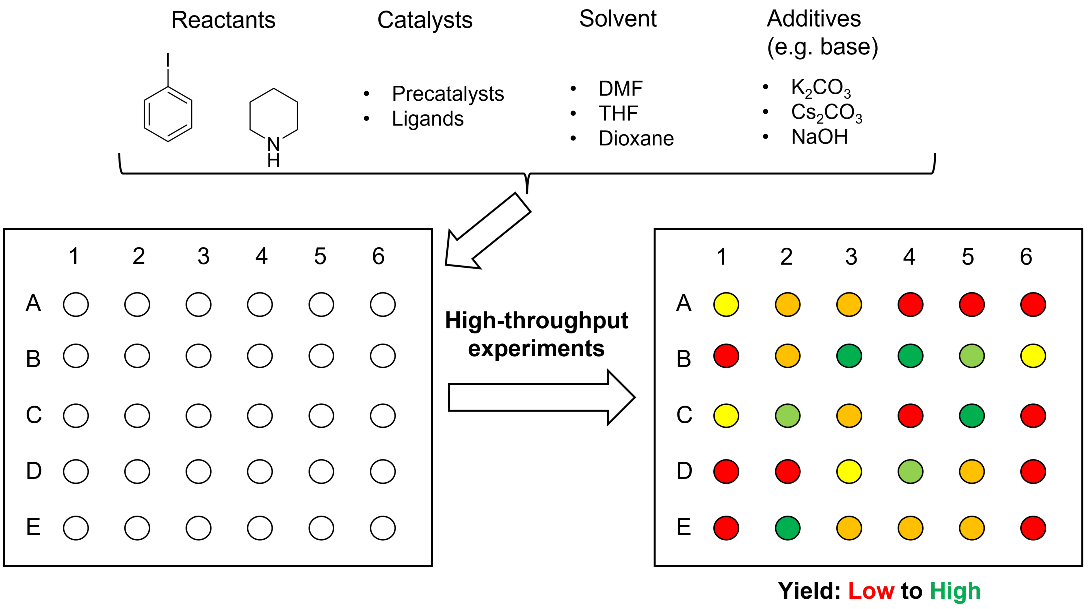


Figure 1.3: Overview of high-throughput experimental screening, using multi-well plates, to determine yields for a set combination of reaction conditions.

Traditional workflows, for example, automated high-throughput experimental screening, could be considered to be much faster as it allows for hundreds of reactions to be carried out in a day by one chemist, in a system that is closely related to the final target system. However, this approach requires a large library of physical catalysts which must either be synthesised or curated from commercial sources. Commercial catalyst screening kits are available from chemical vendors such as Sigma-Aldrich. The ligands available in these kits are limited to those that have been made synthetically or studied previously, introducing the potential for a biased sampling of chemical space.

For more bespoke chemical transformations, highly specialised ligands are often required which are not present in most physical libraries. These unique transformations are often identified through mechanistic insight and can therefore be studied computationally in a comparatively short time frame due to the fewer mechanistic pathways that need to be explored. Several examples of catalysts identified via this approach are already available.[10–13] For systems which are completely unconstrained with a lack of fundamental understanding, containing numerous possible mechanistic pathways, a different approach is required. A combination of both experimental and computational methods, where the initial screening is carried out experimentally, optimised computationally, and then repeated in an iterative process could be a potential solution. Another alternative is to use a purely structural approach where mechanistic considerations are excluded entirely and a vast amount of curated experimental data is used to link chemical structure to performance metrics such as reaction rate or yield.

Current exploration of ligand space is done synthetically, usually starting from a lead compound which has been shown to have some activity experimentally. Analogues are then generated using chemical intuition and tested. Such an approach is viable for reactions with a high mechanistic understanding and/or previous ligand knowledge. However, for catalysts with little mechanistic exploration or limited to no ligand understanding such as those for base metals or for new chemical transformations, this approach becomes less viable. In this case, a lead ligand is likely to be identified by high-throughput experimental screening of a select few ligands, again chosen through chemical intuition. In some cases, this approach must also have sound theoretical backing before any experiments are approved, which due to the nature of the reaction of interest is difficult. Only through trial and error and growing expertise do success rates increase. Moving from experimental to computational screening offers multiple benefits in these cases by opening up the available ligand structures to community or proprietary 3D structure libraries and reducing the experimental resources required.

Computationally driven catalyst design is an exciting prospect that will allow us to think outside of our current understanding and preconceptions built up over centuries of chemical research. Starting from large curated databases of three-dimensional chemical structures is a good starting point for ensuring that the process is both chemically diverse and synthetically viable (**Figure 1.4**). Identifying catalysts that are impractical to make is a waste of resources but confining to only the synthetically proven risks the exclusion of potentially promising candidates. A careful balance must be made between chemical diversity and synthetic availability.

The exploration of new ligands is limited to the availability of a lead ligand whereby its functionalisation or modification is explored. Where no suitable starting point is available significant resources are required to try and identify a lead compound. For poorly understood catalytic systems such as first-row transition metal catalysts, the chance of success is extremely low and in the majority of cases does not justify the resources required to explore. Using chemical structures from a structural database significantly reduces the raw resources, as only a computer is required. Structures contained in databases such as the Cambridge Structural Database (CSD) have 3D X-ray data and therefore, have been proven to be synthetically viable with a well-described synthetic pathway. The 3D structural data can be used to screen for potential activity computationally before any commitment is made to any experimental exploration. The development of tools that can aid the process of identification and prediction of activity from the CSD can significantly speed up the identification and exploration of novel ligands and catalytic systems.

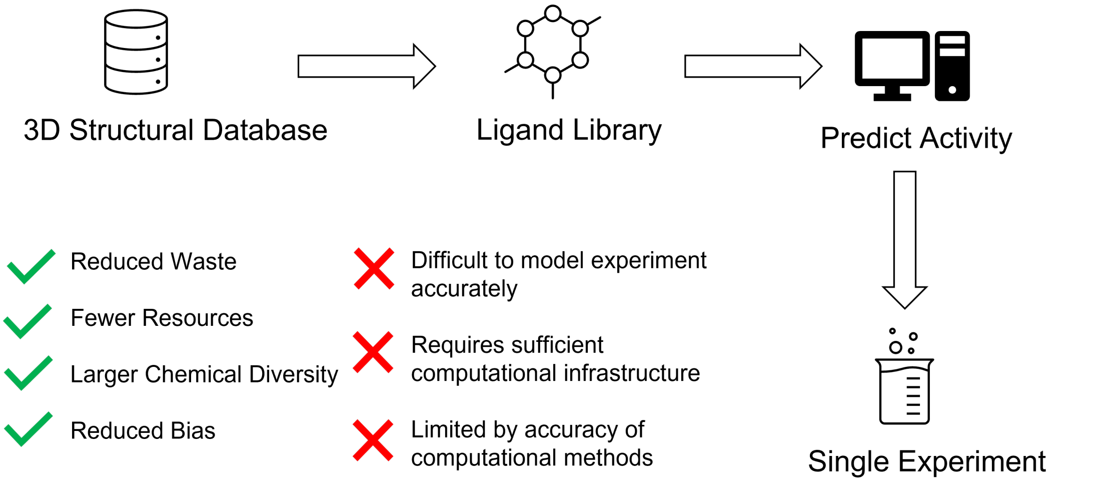

3D Structural Database
Ligand Library
Predict Activity
Single Experiment
✓ Reduced Waste
✓ Fewer Resources
✓ Larger Chemical Diversity
✓ Reduced Bias
✗ Difficult to model experiment accurately
✗ Requires sufficient computational infrastructure
✗ Limited by accuracy of computational methods

Figure 1.4: Ideal workflow for ligand/condition identification using computational screening from a 3D structure database to predict activity. Resulting in a single experiment to validate the prediction.

The development of predictive methodologies that can be applied systematically and that can be implemented via automation are a useful complementary tool that can be used without exhausting human resources. Automation can take care of the monotonous, tedious and error-prone tasks of a systematic study, while humans make the intellectual decisions to guide the study. The growing stores of chemical structural data available in databases can be exploited by machines in a manner that is more precise and objective without introducing bias from a chemist's preconceptions. Such an approach could allow molecular design to go beyond the limitations of a traditional, human-driven approach.

An ideal scenario would consist of an entirely computational screening process leading up to a single experiment. Predictions of which would come from either directly calculated properties or machine learning models. The most effective approach is likely to be a combination of both experimental and computational methods, harnessing the power of advancements in synthesis, physical chemistry and theoretical methods in order to study synthetic pathways, the effect of catalyst structure and how it affects catalytic activity. As technology advances the role of computational methods will become more pronounced. Therefore, the development of user-friendly tools accessible to the wider synthetic community would be extremely valuable.

1.1.2 Sustainability and Assessment of Catalysts in Industry

The most common metals used in organometallic catalysis are precious metals (second and third rows of the periodic table), namely rhodium, platinum, ruthenium and palladium. Precious metals are commonly used as catalysts due to several unique characteristics:

- **Stability:** Precious metals are resistant to corrosion and oxidation with many of their low-valent complexes being stable enough to oxygen that they can be used under ambient conditions, without the need for an inert atmosphere. Precious metal complexes also exhibit low ligand lability making them extremely stable in solution.[14]
- **Oxidation state changes:** A large number of important synthetically relevant chemical reactions proceed via an oxidative addition-reductive elimination mechanism which requires two-electron oxidation state changes at the metal centre, which precious metals readily undergo.
- **Pi bond acidity:** Precious metal complexes often exhibit a high affinity for pi bonds, present in many functional groups in organic molecules, making them well-suited for reactions which proceed via pi-bond activation.
- **Selectivity:** Precious metals often display unique selectivity for specific transformations. For example, the oxidation of ethylene with silver, palladium or platinum yields ethylene oxide, acetaldehyde or CO and H_2O respectively.[15]
- **Characterisation:** Due to their stability, precious metal complexes are often easily isolated and characterised by common techniques such as nuclear magnetic resonance (NMR) and X-ray diffraction (XRD). Precious metals generally form diamagnetic complexes making them easier to analyse by NMR. Therefore, relating structure to activity and selectivity is more straightforward.

Despite the prevalence of precious metal catalysts in the chemical industry, there are several problems relating to their use. By definition precious metals are scarce, and therefore are in low natural abundance, very expensive and susceptible to supply fluctuations. As the global distribution of precious metals in the earth's crust is not uniform, they often have to be imported from other parts of the world. In 2011, the British Geological Survey released the list of metals at risk of supply disruption, with ruthenium, rhodium, palladium, osmium, iridium and platinum being among those at the highest risk.[16] Precious metals also have several environmental implications. Due to their low natural abundance, the mining of these metals requires increased use of fossil fuels and CO_2 emissions. Furthermore, it is predicted that less than 1% of precious metals are recycled due to the economic viability of these processes. It is, therefore, of high importance that we move away from precious metals to more sustainable alternatives.

The use of metals with a high abundance and balanced global distribution is therefore an attractive alternative. The first-row transition metals (base metals) offer additional advantages such as low cost and global availability. Base metals are considered to be metals with minimal safety concerns, except nickel and chromium, compared to often toxic precious metals. For example, 1300 ppm of iron is tolerable in active pharmaceutical ingredients compared to 10 ppm for palladium, which also requires special measures to remove, creating large amounts of waste.[17] While first-row transition metals could provide a sustainable alternative there are several challenges with their use in catalysis due to due reactivity creating problems with stability, selectivity and scope:

- **Stability**: Base metal complexes are often sensitive to oxidising conditions and require an inert atmosphere to prevent degradation. However, in some cases, this can be avoided by the generation of the active species in situ from a suitable precursor.
- **Oxidation state changes**: Unlike precious metals which undergo the required two electron change in the oxidation state required for oxidative addition-reductive elimination processes, base metals generally undergo single electron transfer processes. This means that special methods have to be found to promote two-electron processes.
- **Selectivity**: Base metals favour single electron transfer, this can lead to the generation of radical species which are hard to control and generate products in an unselective manner.[18] Ultimately this leads to a lower functional group tolerance and limited substrate scope.

- **Characterisation**: The lower stability of base metal complexes means that they are difficult to isolate and characterise. Complexes are often paramagnetic due to the presence of a weak ligand field making NMR analysis of these species challenging.

Due to the preference for paramagnetic complexes, many studies on base metal catalysts have been conducted on low-spin, diamagnetic carbonyl and cyanide complexes. Recently there has been renewed interest in paramagnetic complexes for iron, nickel and cobalt for catalysis. A base metal-catalysed C−C coupling reaction was recently discovered where the only ligands available in the reaction were weak-field halides, *O*–donors, solvent or additives. Assuming the resulting catalytic species is high-spin due to the weak ligand field available, this implies that paramagnetic base metal complexes are active catalysts.[17] In fact many recent high-spin first-row metal reactions have excellent selectivity.[14] However, it is not understood for which cases a high-spin catalyst would be advantageous. This is due to a lack of fundamental understanding of the reaction mechanisms and coordinating environments of the metals.

Table 1.1: Comparison of precious metals and base metals in catalysis.

Property	Precious Metals	Base Metals
Stability	• Relatively stable to oxygen • Low ligand lability	• Sensitive to oxidising conditions • Labile ligands
Oxidation State Changes	• Readily undergo 2-electron oxidation state changes	• Generally undergo 1-electron oxidation state changes
Selectivity	• Display unique selectivity • High pi bond affinity	• Unselective radical species are common • Lower functional group tolerance
Characterisation	• Stable complexes • Usually diamagnetic	• Low stability • Paramagnetic complexes are common

The creation of new complexes of base metals that exhibit greater stability, selectivity and have a greater substrate scope are required to move away from precious metals to a more sustainable and cheaper catalysis space. As ligands in these reactions are often understudied, using 3D structure databases as a source of potential ligands is an attractive approach to identifying novel ligands. Experimental approaches in these cases are often trial and error and resource-demanding due to the lack of a suitable lead structure. Using a computational approach both saves raw resources and provides a larger more diverse coverage of chemical space, potentially increasing the chance of finding a suitable lead ligand.

1.2 Computational Approaches to Homogeneous Catalyst Design

This brief introduction will cover a range of computational methods, software and approaches to the computational design of homogeneous catalysts. For a more extensive review of the area, the following reviews are recommended. Li and Merz for a review of metal ion bonding using molecular modelling.[19] Pidko et al. for computational approaches to transition metal catalysis.[20] Fey et al. for the use of ligand descriptors for the prediction of the chemical properties of transition metal complexes.[21] Jensen et al. for a review of the challenges faced in in silico catalyst design and Kulik et al. for an overview of the applications of high-throughput screening and machine learning to inorganic catalyst discovery and the prediction of chemical properties.[22,23]

The advancement of computational modelling methods along with improved computational power has reached a point where it is a useful complementary tool in catalysis for the interpretation of experimental results, either by predicting activity or selectivity, or by guiding experimental workflows. Predictive strategies in catalyst design can be divided into three main groups: (1) trial and error, (2) prediction models and (3) automated design.[22]

Trial and error methods cover the use of interactive computational tools along with chemical intuition to test ideas through the use of 3D modelling or simulations to guide molecular design. Examples range from using 3D molecular models to represent the volume and shape of the catalytic site to the calculation of free energy profiles along the reaction pathway.[24,25] Where multi-step reactions are studied, computational cost can be high, especially where high accuracy is required.

Prediction models aim to relate statistical data with quantitative or qualitative structure-activity/property relationships (QSAR/QSPR) through the use of a set of ligand descriptors to correlate ligand properties to a desired catalyst property, such as activity or selectivity. Predictive models can quickly predict the properties of novel compounds similar to those used in the training data. However, these models only relate to a specific region of chemical space and therefore, predictions outside this space are unreliable.

Automated design is an umbrella term that covers the automation of the computational tasks associated with the identification of catalyst candidates. Similar to predictive models, automated design uses an automated predictive model to navigate chemical space to desired property regions and generate candidate molecules without input from the user. Automated methods make use of the ever-growing library of chemical knowledge and exploit it in a way that humans are incapable of. Another benefit of automated design is the removal of all bias introduced by humans, allowing molecular design to go beyond human-imposed limitations.

The following review covers some of the most useful design methods relating to this project, which focus on predictive modelling in homogeneous catalysis. For a more extensive review of catalyst design methods see a recent review paper by Foscato and Jensen.[22]

Molecules are incredibly diverse; ranging from small simple organic molecules with several hundred Daltons in molecular weight, organometallic complexes containing both metallic and non-metallic elements, to large complex materials weighing hundreds of thousands of Daltons. As molecules are three-dimensional by nature; effectively representing them in a manner which captures their functionality, diversity and orientation in three-dimensional space is a major challenge.

In catalyst design, the intrinsic properties are often described by descriptors, which encapsulate both the steric and electronic properties of the catalyst. A large array of descriptors exist in the literature, most of which have been developed for drug design.[22] However, descriptors are also being developed for applications to catalysts such as those describing metal-ligand bonds and ligand steric properties.[21] An example of such descriptors from Fey and co-workers aims to describe both phosphine and carbene ligands and their use in transition metal catalysis.[26,27] Examples of typical ligand descriptors include atomic charges, pK_a values, HOMO-LUMO gaps, geometrical features (sterimol parameters, bond angles/distances) and the Tolman cone angle. In cases where steric interactions are dominant such as stereoselective reactions, molecular interactions fields (MIFs) have been used to predict enantiomeric excesses and to identify regions of maximum stereochemical induction.[28,29] Noncovalent interactions should also be considered, as they can significantly affect catalyst efficiency and selectivity.[30] Recently several descriptors have been developed to take these interactions into account.[31] Fey and Durand provide an extensive view of the current literature surrounding the use of ligand descriptors in catalyst design.[21]

The most commonly used approach for computational catalyst design is ligand additivity, whereby the properties of heteroleptic complexes can be inferred from combinations of homoleptic complexes.[32] Additivity is also used in fragmentation methods where a ligand is divided into smaller groups, with the properties of these smaller components used to predict to properties of the combined structure (**Figure 1.5**).[33] Fragmentation can be used to either predict the properties of a singular ligand or a transition metal complex using the properties of the individual ligands.

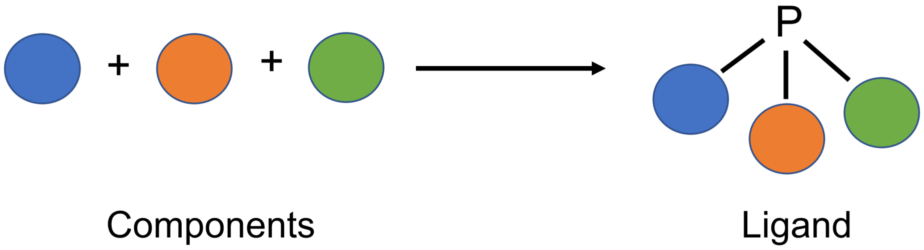


Figure 1.5: Example of ligand additivity where the properties of components are used to predict the properties of a ligand.

1.2.1 Ligand Knowledge Bases

Traditional methods of screening catalysts include experimental optimisation and high throughput screening. The reproducibility and time/cost restrictions of screening large numbers of ligands experimentally make it difficult to compare ligand performance over a wide range of substrates and reaction conditions. 50 years ago Tolman introduced experimentally measured descriptors to rationalise the properties of phosphorous ligands.[34] Molecular descriptors allow the mapping of chemical space and can be used as a tool to identify systematic trends in reactivity and stability. The emergence of quantum mechanical methods allowed for these descriptors to be calculated computationally. Fey et al. have developed computational methods of screening large numbers of free and complexed ligands through the combination of structural databases and density functional theory calculations to create a database of structural and electronic descriptors.[26,27] Descriptors from experimental and computational studies are used to create a 'ligand knowledge base'

(LKB) which aims to describe all of the available chemical space. Currently, databases exist for monodentate and bidentate phosphorus ligands, carbenes and recently small organic bidentate ligands.[35–37] Curation of descriptors into a database provides a data-rich environment which can be drawn upon for use in data analysis methods or for the identification of similar compounds.

Descriptors contained in these ligand knowledge bases are limited as they need to account for all transition metals. Descriptors such as bite angle are sensitive to the metal, its oxidation state and its electronic configuration. The metal the descriptor is generated from imposes a structural demand on the ligand affecting the geometry and ligand field stabilisation.[21] Therefore, descriptors used may not correctly describe the correct environment for certain metals and reactions with different coordination environments than the one the descriptor was calculated from.

Ligand knowledge bases offer an alternative method for ligand design, that is both potentially faster and applicable to a wider chemical space. Interpretation of large datasets to predict chemical properties enables a more direct route for catalyst design rather than screening a wide chemical space. However, they are still underdeveloped.

Recently, Gensch et al. developed Kraken, a virtual open-access library for monodentate phosphorous ligands.[33] Semiempirical and DFT descriptors were calculated for a library of 1558 organophosphorus compounds. Principal component analysis was used to map the chemical space to allow for the identification of suitable ligands for a specific problem. Furthermore, a fragment-based approach was used to generate a library of 300,000 computationally generated ligands and machine learning was used to predict their chemical properties. Inverse ligand design was used to successfully identify a set of ligands for enantioselective Pd-catalysed $sp^3{-}sp^2$ cross-coupling between alkylboronic acids and aryl halides. This approach required independent experimental studies to identify key chemical features for high activity in order to identify the correct ligand space. While a fragment-based approach is able to produce and map a large chemical space the synthetic viability of the ligands generated from this approach is questionable. Chemical space mapping can bridge the gap between computational and synthetic chemists due to the easily interpretable data, allowing synthetic chemists to perform computer-assisted interactive ligand exploration without an in-depth knowledge of computational chemistry.

1.2.2 Quantum Guided Molecular Mechanics (Q2MM)

Asymmetric catalysis is an essential method for preparing biologically active compounds. Its importance has led to the computational design of such catalysts being described as a 'holy grail' of chemistry.[38] Traditional methods involve high-throughput experimental screening, conducting hundreds of reactions at once in multi-well plates, covering a large range of reaction conditions. Determination of activity and selectivity is achieved through chromatography. It is often very expensive to cover a wide range of chemical space and reaction conditions, involving personnel, time, equipment and resource acquisition. Reducing the number of ligands screened by using computational approaches will lead to a significant reduction in both time and cost.

Computational quantum mechanical methods have long been used for the identification of transition states to rationalize and predict stereochemical outcomes. Due to the large computational cost of quantum mechanical methods such as DFT, especially for determining transition state structures, only a small number of conformations are sampled and the screening of large ligand libraries is unfeasible compared to the time required for high-throughput screening experiments. Several methods have been developed to make stereoselective screening computationally viable.

The AARON toolkit generates small libraries of transition states for a set of ligands using a semiempirical conformational search. The structures of the lowest energy conformations are then optimised using DFT to generate predictions.[39] Sigman and co-workers used a different but complementary approach whereby the stereoselectivity of a small training set is fitted against a set of physicochemical parameters.[40] This approach is based upon the steric and electronic effects of substituents on the ligand instead of mechanistic information and can therefore only be used on closely related structures.

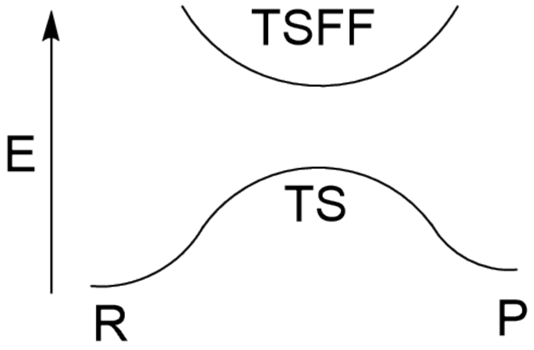


Figure 1.6: Comparison of Q2MM (TSFF) and traditional (TS) transition state modelling methods.

The quantum-guided molecular mechanics method (Q2MM), uses DFT-generated transition states to parameterise a transition state force field (TSFF). The TSFF is built upon a standard force field such as AMBER or MM3 and parameters are added from the DFT training data to parameterise metals and any uncommon bonding motifs, present in transition states but not present in the standard force field. Due to the inability of molecular mechanics methods to optimise to saddle points, the transition state is treated as a minimum (**Figure 1.6**). This force field is used to calculate the structure of the transition state. A subsequent conformational search generates a set of conformers for each pathway, which is repeated for the list of ligands and substrates. Stereoselectivity is then determined through the Boltzmann averaged energy of the conformational ensemble. Q2MM has been applied to several reaction types such as osmium-catalysed dihydroxylation, rhodium-catalysed hydrogenations and stereoselective additions to aldehydes.[41–43] Experimental stereoselectivities were reproduced with a mean absolute error of 2.8 kJmol^{-1}, where a prediction of 99% *e.e.* would yield 97-99.7% in the laboratory.

For a typical catalyst-substrate combination with less than 10 rotatable bonds, all conformations can be generated in about one hour on one CPU, offering significant speed increases compared to the DFT equivalent.[44] Generally most Q2MM software suffers from poor usability and time inefficiencies, however, the Wiest/Norrby and Moitessier groups have developed CatVS and ACE to begin to address this.[44,45]

1.2.3 Machine Learning Approaches to Catalyst Design

Machine learning can be used to both predict a property of a chemical of interest and/or to identify key properties for a desired output variable. For example, in catalyst design, a model could be used to predict the activity of the catalyst by the prediction of the activation energy of a specific reaction or use a predetermined activation energy value to identify the key properties of the catalyst required for high activity. The general cheminformatic approach is shown in **Figure 1.7**. Under this regime, machine learning models are built by gathering/generating and curating a large quantity of data, encoding the molecules as inputs (descriptors/features) and mapping them to the desired output property.

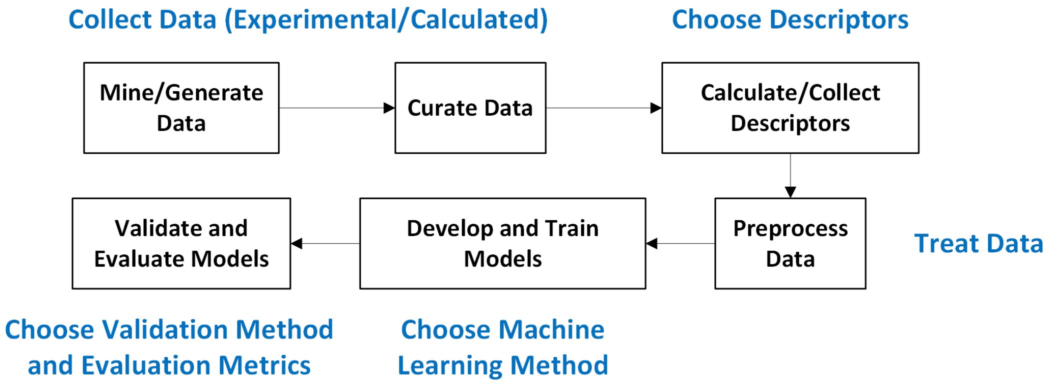

Collect Data (Experimental/Calculated)
Choose Descriptors
Mine/Generate Data
Curate Data
Calculate/Collect Descriptors
Preprocess Data
Treat Data
Develop and Train Models
Validate and Evaluate Models
Choose Validation Method and Evaluation Metrics
Choose Machine Learning Method

Figure 1.7: A general machine learning workflow for building a cheminformatics model.

Machine learning is a statistical-based method that can construct powerful correlation or classification models from single or multi-variable datasets in order to predict outcomes. The majority of machine learning methods are based on the assumption that a relationship exists between intrinsic properties such as atomic properties, and a global property, such as catalytic activity.[46] Often a simple linear relationship is sufficient to correlate properties but more complex nonlinear methods may be required depending on the complexity of the system of interest. However, correlation does not always imply causation and therefore, the applicability of these models should be carefully evaluated. Artificial neural networks and kernel ridge regression models are the most commonly used machine learning methods in homogeneous catalyst design.[47]

The selection of suitable descriptors is a key factor influencing the predictive power of the model and can be guided by either mechanistic or structural insights. However, these are not necessary as one of the advantages of machine learning models is that they can be used when the reaction mechanism is unknown.[48]

Deep learning uses a method of learning in which simple internal representations are combined to form complex objects. Chemistry is a prime example of a field that in principle is perfect for deep learning. The behaviour of molecules can not be determined by atoms alone, but by their grouping into functional groups and the interactions between functional groups at differing ranges. However, molecules also provide a set of challenges for use with machine learning. Machine learning is the most successful in fields where there is an abundance of data, with the most useful having datasets in the millions and even billions of data points. In chemistry, lab-derived data is the gold standard, however, currently available databases do not possess data points within the same order of magnitude.[49] Therefore, computational chemistry has emerged as the leading source of data due to the ability to generate a large volume of data in a much shorter time than it would take experimentally in the laboratory. However, the accuracy of these results is poorer than those collected experimentally.

Another common machine learning method, neural networks, has also not been used extensively due to the large amount of data required to train the models. The lack of curated, accurate, consistent data, along with the multimolecular nature of catalytic processes are major challenges in the application of these methods.[22] A recent example of the application of neural networks by Denmark and co-workers used a neural network to predict the stereoselectivity of the addition of thiols to imines.[50] It was found that even using training data with low selectivity (<80% e.e.) highly selective catalysts could still be predicted well outside of the training set. Recent applications by the Kulik group have predicted redox potentials, spin-state splitting and atomisation energies of organometallic complexes.[51–54] A notable example is the identification of metal complexes for the selective oxidation of methane to methanol.[55] Kulik et al. used a fragment-based approach with an artificial neural network to screen 16M macrocycles. They found that low-spin Fe(II) complexes with strong-field (P, S- coordinating) ligands gave the best balance between hydrogen atom transfer and methanol release from the metal. They also found that high valence metals were rate limited by slow methanol release and that negatively charged axial ligands were critical to promoting the release of methanol from the metal centre. Mn(II) and low-spin

Fe(II) catalysts were predicted to have high turnovers. However, they were not tested experimentally.

Accurate virtual high-throughput screening of transition metal complexes is challenging due to the possibility of multireference (MR) character.[56] Complexes with high multireference character complicates the calculation and prediction of chemical properties. Kulik et al. recently developed a method for the prediction of multireference character using a neural network.[56] They computed MR diagnostics for 5,000 ligands in the CSD and found that MR character correlated linearly with the inverse value of the average bond order over the entire molecule. They also observed that MR character can be inferred from the sum of the MR character of the ligands. Therefore, ligand additivity was used to train a neural network to predict MR character using the properties of the constituent ligands. This work is only applicable to equilibrium structures, therefore not tested on structures such as transition states. The ability to predict MR character without expensive computational approaches will be important in making high-throughput computational screening viable for a large range of transition metal catalysts, especially for metals which favour open-shell species.

Random forest models have been used to predict the performance of palladium-based catalysts used in amination reactions.[57] Data was collected from over 4000 high throughput experiments with a set of 120 atomic, molecular and vibrational descriptors used to model the catalyst, ligands, substrates and additives. The random forest model showed superior performance over linear regression analysis for predicting the tolerance of the palladium catalyst for isoxazoles during C−N bond formation. Recently random forest models were also applied to the asymmetric relay Heck reaction to predict stereoselectivity.[58] Quantum chemically generated organic descriptors were found to predict enantiomeric excess well, with an RMSE of 8.0±1.3%. Ligands were generated using a fragment-based approach by varying R groups of common ligand scaffolds.

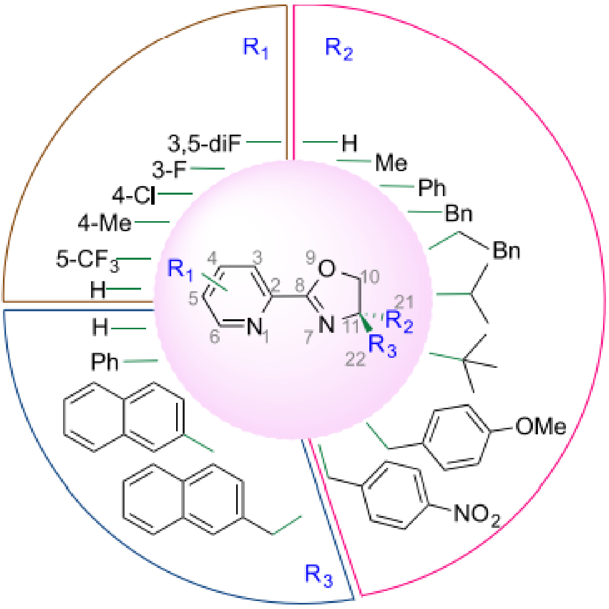


Figure 1.8: Fragmentation approach explored by Das et al. for the relay Heck reaction.[58]

Gaussian process (GP) models can be used to predict properties from unseen inputs. Predictions are in the form of a mean value and associated variance relating to the confidence of the prediction. The variance can be used to judge the reliability of the prediction and whether it should be discarded and explicitly calculated and used to retrain the model. This systematic approach allows retraining of the model increasing the quality of predictions and allowing for increased refinement. Gaussian processes based prediction models are the most attractive machine learning method available, due to being able to build models with sparse data from small to medium-sized datasets from multiple sources and differing qualities.[59] However, they have not yet been applied to homogeneous catalysis, but initial studies of heterogeneous catalysts look promising.[59]

The use of ligand knowledge bases as a source of descriptors for machine learning methods shows promise for the high-throughput identification of ligands for transition metal catalysts. The first major milestone for deep learning in chemistry came from Dahl et al. in 2012 who won the ‘Merck Molecular Activity Challenge’ for their multitask neural network which was able to predict the molecular activity of molecules in 15 different sites in the body with greater accuracy than traditional machine learning methods such as boosted decision trees.[60] However, the first breakthrough for the modelling of molecules was the ‘molecular auto-encoder’ by Gómez-Bombarelli et al..[61] Since then there has been a large increase in advancements in the modelling of molecules. The most notable of which involves the representation of molecular structures. In their initial paper, Gómez-Bombarelli et al. used the SMILES representation, however, the main disadvantage of SMILES for molecular representation is the fact that it is not unique. Therefore, the introduction of graph-based representations attempted to solve this as well as provide a way to represent three-dimensional structures. There are now many different representations available, meaning that there is a lack of standards for both benchmarking and comparing different approaches. The are very few examples of machine learning successfully predicting ligand activity which have been validated experimentally.[57]

The lack of 'failed' experimental data available in both the literature and chemical databases makes building effective models from widely available data sources extremely difficult. While it is important to know what works well, it is also important to know what doesn't. The standard of only publishing 'good' results and not the 'failed' data is counter-productive and makes the curation of datasets for machine learning challenging. Making the publication of all data along with a scientific publication will aid the uptake and quality of future machine learning models.

1.3 Choice of Model Reaction: The Ullmann-Goldberg Reaction

Palladium has become one of the most important transition metals used in catalysis, with applications in cross-coupling, insertions and other important chemical transformations. It is now the most commonly used transition metal catalyst in the pharmaceutical industry, being used in important synthetic reactions such as Suzuki-Miyaura, Buchwald-Hartwig and Pd-borylation.[62] However, as palladium is a precious metal it has a low abundance and is, therefore, very expensive and avoided where possible in large-scale chemical production. A cheaper and more sustainable metal catalyst is required in the future.

1.3.1 The Importance of C-N Bond Formation In Synthetic Chemistry

The addition of nitrogen to sp^2 carbon centres is an important chemical transformation in organic synthesis. A significant number of pharmaceutical compounds contain aryl carbon-nitrogen bonds (**Figure 1.9**). Common synthetic routes to generate aryl amines are shown in **Figure 1.10**. Route A shows the introduction of a C−N bond via nitration of the benzene ring at a C−H bond followed by reduction of the nitro group to the amine. The amine can then be functionalised to introduce R groups. Functional group tolerance for route A is limited due to the harsh conditions required to introduce the nitro group. Nucleophilic aromatic substitution (S_NAr) can also be used to form arylamines (Route B). Although the reaction conditions are mild, strong electron-withdrawing groups are required at the *ortho*- and *para*- positions to activate the C−X bond. Route C is the superior route as it requires milder reaction conditions and has a wider substrate scope. Halogenation of the benzene ring is followed by transition metal-catalysed conversion of the C−X bond to a C−N bond.

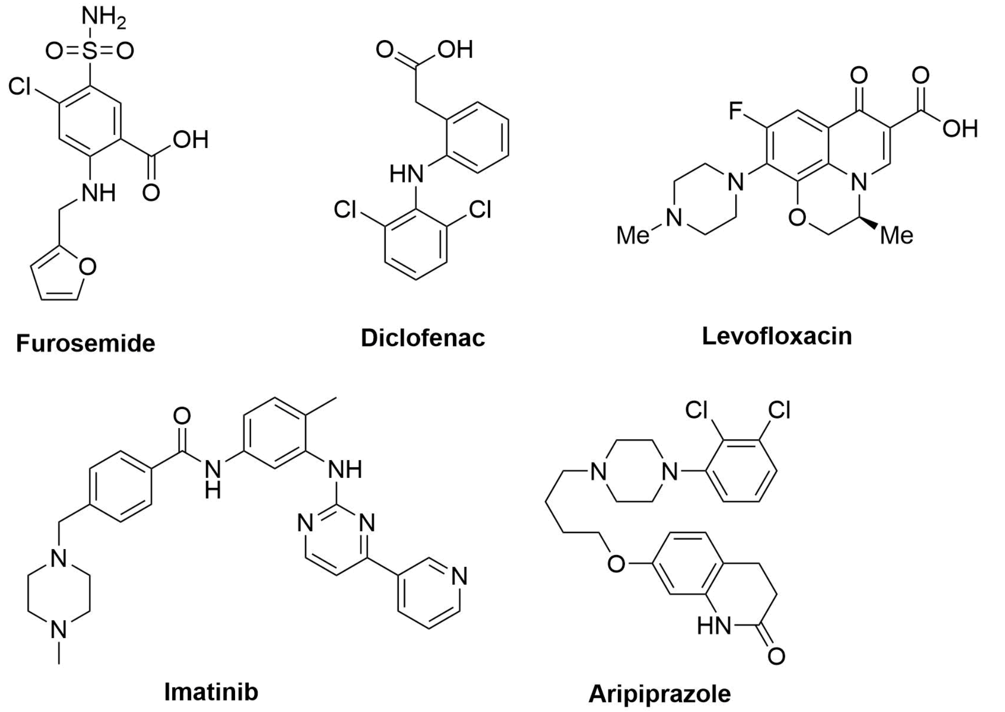


Figure 1.9: Examples of pharmaceuticals that contain aryl carbon-nitrogen bonds.

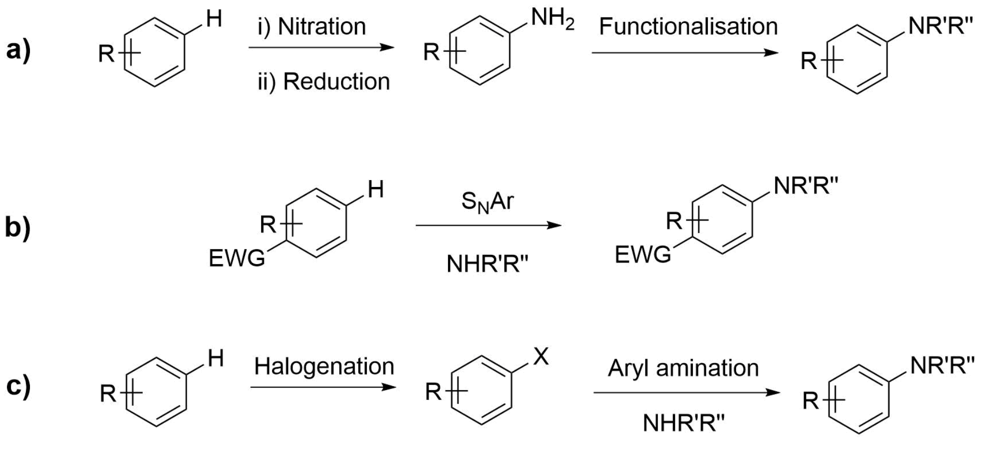
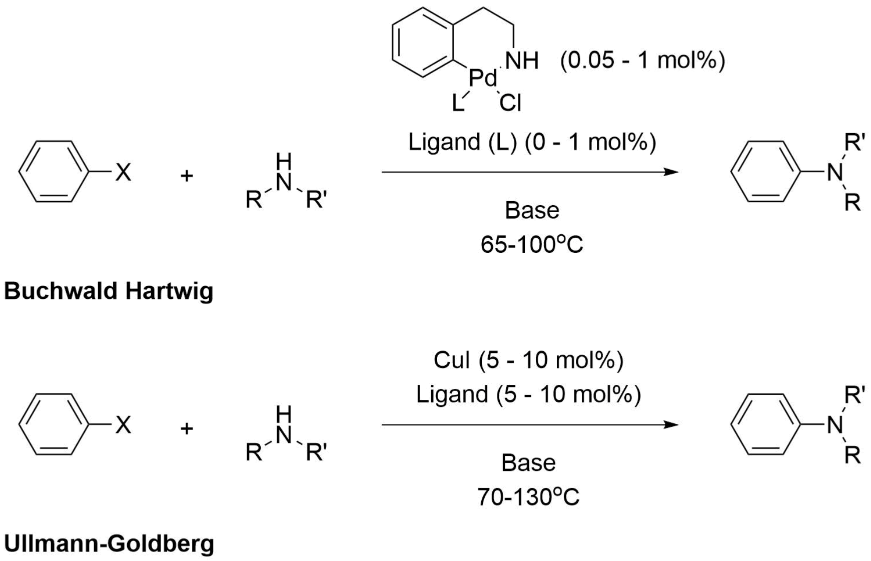


Figure 1.10: General reaction schemes to generate arylamines. X = halide; EWG = electron-withdrawing group.

The palladium-catalysed Buchwald-Hartwig amination is the most common transition metal-catalysed process for converting C−X bonds to C−N bonds. The Buchwald-Hartwig amination uses phosphines as ancillary ligands, the most effective of which being dialkylbiaryl phosphines (e.g. BrettPhos and RuPhos).[63] While palladium-catalysed aminations have been extensively studied, the reactions have some disadvantages such as high cost, the high molecular weight of ligands and the toxicity of palladium.

Copper is a suitable replacement for palladium as it shares similar chemistry. For example, the traditional Ullmann coupling and Ullmann-Goldberg reaction share similar chemistry to palladium cross-coupling and the Buchwald-Hartwig reaction respectively. Currently, 10% of drug discovery and 5% of process chemistry synthetic routes contain an aryl-amine bond-forming step.[62] Process chemistry tends to avoid the use of toxic and expensive catalysts in industrial processes due to cost and heavy metal contamination. Introducing copper as a replacement for palladium will reduce the cost and environmental impact of industrial processes as well as reduce supply issues due to its higher natural abundance.

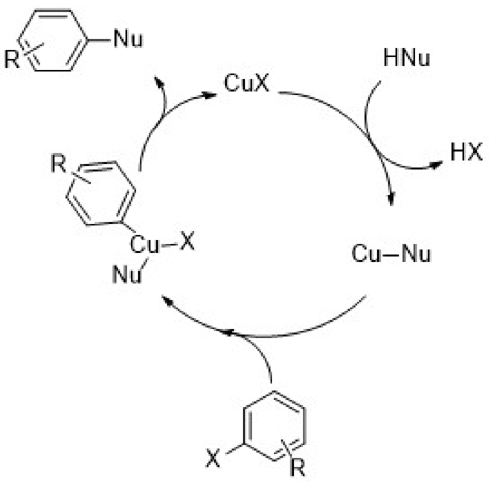
However, copper has several issues making it less efficient than palladium for cross-coupling reactions. First of all its complexes have low stability due to the ligands being highly labile. Secondly, as the ligand and the nucleophile are N/O donors, high ligand and catalyst loading are required to achieve a suitable population of the active catalytic species in solution. Finally, the understanding of ligand properties is not very well understood and therefore, designing ligands for specific chemical transformations is difficult. This makes the Ullmann-Goldberg reaction an attractive target for the application of high-throughput screening and machine-learning approaches.



Scheme 1: General reaction schemes for the palladium-catalysed Buchwald-Hartwig amination and Ullmann-Goldberg reaction.

1.3.2 Copper-catalysed Cross-Coupling

In 1901 Fritz Ullmann reported that copper compounds were able to catalyse the formation of biaryl moieties through the coupling of two molecules of an aryl halide.[64] This reaction is now often referred to as the 'classical Ullmann reaction'. The generally agreed mechanism for this reaction involves the insertion of the copper into the aryl-halide bond, which then undergoes oxidative addition with a second molecule of the aryl-halide and subsequently eliminates the product through reductive elimination.



Scheme 2: General reaction scheme for the Ullmann-Goldberg reaction.

In 1903 Ullmann applied the same methodology for the synthesis of N-aryl amines, using stoichiometric quantities of copper, and to ethers in 1905, commonly referred to as 'Ullmann condensation'.[65,66] Irma Goldberg, in 1906, reported the first copper-catalysed synthesis of aryl amides and also improved upon Ullmann's original N-aryl amine synthesis by moving from a stoichiometric to catalytic amount of copper (**Scheme 2**).[67] Later in 1929, William Hurtley reported the coupling between *o*-bromobenzoic acid and β-dicarbonyls mediated by either Cu bronze or $Cu(OAc)_2$.[68] These early reactions, however, required harsh conditions such as high temperature, stoichiometric copper, strong bases, long reaction times and high boiling point polar solvents. There was also a limited substrate scope, generally requiring electron-poor aromatic substrates. The use of stoichiometric copper was required due to issues relating to the poor solubility of many of the copper sources used.[69]

$$\mathrm{Ar{-}X + Ar{-}X} \xrightarrow{[\mathrm{Cu}]} \mathrm{Ar{-}Ar} \qquad \text{Ullmann reaction 1901} \tag{1.1}$$

$$\mathrm{Ar{-}X + Ar{-}NH_2} \xrightarrow{[\mathrm{Cu}]} \mathrm{Ar{-}NH{-}Ar} \qquad \text{Ullmann-Goldberg reaction 1903} \tag{1.2}$$

$$\mathrm{Ar{-}X + Ar{-}OH} \xrightarrow{[\mathrm{Cu}]} \mathrm{Ar{-}O{-}Ar} \qquad \text{Ullmann-Goldberg reaction 1904} \tag{1.3}$$

$$\mathrm{Ar{-}X + Ar{-}CONH_2} \xrightarrow{[\mathrm{Cu}]} \mathrm{Ar{-}CONH{-}AR} \qquad \text{Goldberg reaction 1906} \tag{1.4}$$

$$\mathrm{Ar{-}X + H{\equiv}R} \xrightarrow{[\mathrm{Cu}]} \mathrm{Ar{\equiv}R} \qquad \text{Castro-Stephens reaction 1963} \tag{1.5}$$

Scheme 3: Notable copper-mediated cross-coupling reactions.

1.3.3 Introduction to Copper

Copper is a period 4 transition metal (base metal) in group 11. Copper has an electronic configuration of $[Ar]4s^1 3d^{10}$ in its elemental form, commonly forming compounds in its +1 and +2 oxidation states. Compounds containing copper(III) are known but are far less common due to the high third ionisation energy of copper.[70] The electronic configuration adopted by copper is energetically favourable compared to $[Ar]4s^2 3d^9$ due to the filled 3d subshell. Copper(I) complexes are diamagnetic with common geometries of two-coordinate linear, three-coordinate trigonal planar, and four-coordinate tetrahedral, whereas, copper(II) is paramagnetic and usually found in a tetragonal coordination environment, with four short equatorial bonds and another one or two longer axial bonds.[71] Other geometries of copper(II) complexes are known, including tetrahedral, square planar, and trigonal bipyramidal geometries.[71] Although copper(I) might be predicted to be in a more stable electronic configuration ([Ar]$4s^0 3d^{10}$) compared to copper(II) ([Ar]$4s^0 3d^9$) due to the full d orbitals, copper(I) is thermodynamically unstable in the presence of water (**Scheme 4**).[71]

$$2\,\mathrm{Cu^{I}(aq)} \rightleftharpoons \mathrm{Cu^{II}(aq)} + \mathrm{Cu^{0}(s)} \qquad E^{\theta} = +0.36\mathrm{V}$$

Scheme 4: Disproportionation of copper(I) to copper(II) and copper (0) in water.

Copper has a relatively high ionisation energy, therefore, copper(I) is relatively unstable compared to copper(0) and undergoes spontaneous disproportionation into copper (II) and copper(0). As a result, complexes of copper(I) often require handling under an inert atmosphere to prevent decomposition when in solution. In the solid form copper(I) is often the more stable oxidation state at moderate temperatures.

1.3.4 Role of Ligands

While the ability of some esters and ketones to accelerate the reaction was known since 1964.[72] It was not known how they affected the reaction, the most common theory was that they increased the solubility of the copper catalyst. The first example of an exogenous ligand being used in the Ullmann reaction was in 1997 by Liebeskind and Buchwald, where they used an over-stoichiometric additive in the copper-catalysed coupling of biaryls and aryl ethers.[73,74] Both Liebeskind and Buchwald proposed different roles for their respective additives. Liebeskind suggested that the additive, thiophenecarboxylate, would accelerate the coupling by facilitating the oxidative addition of the aryl halide to the copper. Buchwald however, proposed that the additives naphthoic acid and caesium carbonate enhanced the solubility of the intermediate copper species.[72,75] There were several theories proposed to explain the effect of ligands on the increased reaction rate. Stabilisation of the active Cu(I) species, increasing the solubility of the copper catalyst, prevention of copper aggregation and multiple ligation to the copper were all proposed as possible mechanisms.[76]

In the following years, the first examples of bidentate ligands were published, which proved to be more efficient than the previously used monodentate ligands. The first example of such ligands was by Buchwald a few years after his initial work with additives, where he used phenanthroline and dibenzylideneacetone (dba) as ligands for the coupling of aryl halides and imidazole.[76] It was suggested that bidentate ligands were more effective as they blocked two coordination sites on the copper, therefore, the nucleophile and aryl coupling partner would be closer together, facilitating reductive elimination.[77] The majority of ligands used in the Ullmann-Goldberg reaction are N and O donor ligands as P ligands have been shown to be not very effective.[78]

Major contributors to identifying the role of ligands were the Taillefer and Buchwald groups, each focusing on different aspects of the problem. Taillefer's group focused on a new class of imine-based ligands for the N- and O- arylation Ullmann-Goldberg reactions, whereas Buchwald's group was focused on kinetic studies of the N-arylation of amides.[79–82]

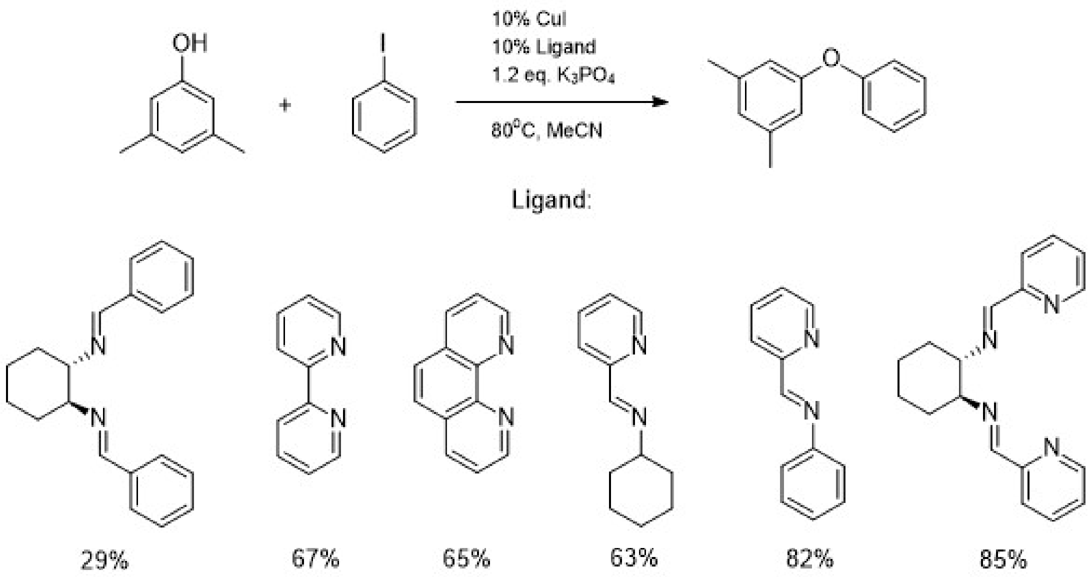
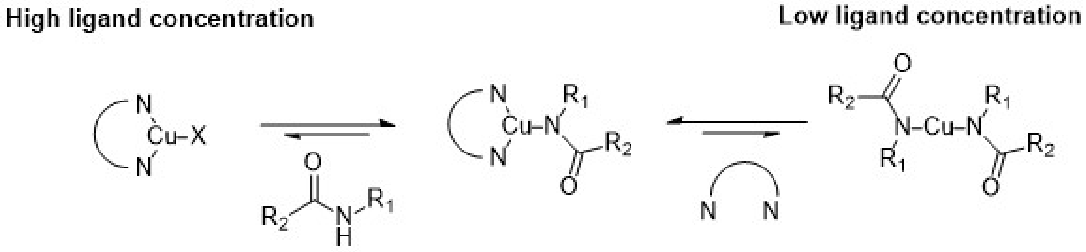


Figure 1.11: Ligands explored by Taillefer et al. and their respective yields.[80]

In 2007, Taillefer's group reported the first ligand structure-activity relationship study. They compared a series of bi-functional iminopyridines with phenanthrolines and bipyridines to determine the effect of ligand structure on catalytic activity (**Figure 1.11**).[80] Ligands containing only imines as the coordinating functional groups were found to be ineffective in increasing the reaction yield. The introduction of a pyridine group onto an aromatic imine showed an increase in yield, with bipyridines and phenanthrolines further enhancing the yield to 82%. Using a tetradentate ligand showed only a slight increase in yield over the more traditional bidentate ligands.[80] Prominent electron effects were also observed, with electron-withdrawing substituents on the imine moiety and electron-donating substituents on the pyridine moiety leading to increased yields. Based on this observation the authors proposed that the imine and pyridine moieties could intervene at different stages in the catalytic cycle. The electron-rich pyridine would donate electron density to the copper atom, increasing its tendency to undergo oxidative addition and the electron-poor imine would decrease electron density on the copper atom, making the catalyst more susceptible to reductive elimination.[80]

Buchwald's group put forward a different point of view based on their kinetic investigations. They proposed that a series of different equilibrium-based processes were involved.[81] When the reaction was conducted at very low concentrations of 0.02 M of catalyst (CuI) it was found that the reaction reached a maximum rate at approx. 0.2 M of ligand. Based on this observation they concluded that the solubilising effect of the ligand was either non-existent or was not the only effect caused by the ligand in the reaction. They, therefore, proposed that the ligand could play a role in preventing the coordination of two amide molecules to the copper atom, making the formation of the mono-amide copper complex, a key reaction intermediate, more favourable, thus increasing the rate of reaction.[81]



Scheme 5: Copper complex equilibrium in presence of high and low ligand concentrations.

Their hypothesis was that at high ligand concentrations, the copper atom is coordinated to the diamine ligand and a halogen (X) from the precatalyst (**Scheme 5**). Ligand substitution between the halogen and the amide anion then occurs to form an intermediate copper amide complex, which goes on to react with the aryl halide to give the coupled product. However, at low ligand concentrations, the copper is coordinated to two molecules of the amide, which undergoes ligand substitution with the diamine ligand to form the same key copper-amide complex, however at a slower rate due to the higher stability of the copper bis-amide complex. The group found that the copper bis-amide species exists as aggregated oligomers in the absence of the diamine ligand, and upon addition of the diamine ligand breaks down into the monomeric species which undergoes fast and mild coupling with the aryl halide to form the coupled product with a reaction half-life $t_{1/2} = 3.1$ min at

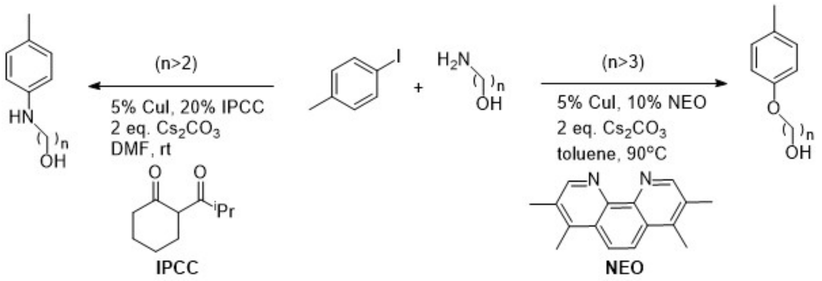
0°C.[81] They also found that at low ligand concentrations, the reaction rate decreases with increasing amide concentration. This demonstrated that the copper bis-amide complex was inactive and that the ligand played an important role in preventing its formation.[81]

The inclusion of ligands in the Ullmann-Goldberg reaction has led to the use of much milder reaction conditions. Temperatures could be reduced to <100°C and copper and ligand loadings reduced from stoichiometric amounts to 5-20 mol%.

1.3.5 Effect of Ligand Properties on Nucleophile Selectivity

Selective arylation of amino-alcohols is an important synthetic transformation, specifically in medicinal chemistry as functionalised amino-alcohols are present in several classes of pharmaceutically active compounds. Advances in copper catalysis have made the Ullmann-Goldberg reaction an attractive method for this type of functional group incorporation. However, the selectivity of either N or O is dependent on several factors, the coordinating ability of the amino-alcohol and the nature of the ligand.

The first breakthrough in the understanding of N/O selectivity came from Buchwald et al. in 2002.[83] In their study of the arylation of β-amino-alcohols, it was found that changing the reaction conditions had a dramatic effect on the selectivity of the reaction. Using NaOH as a base and DMSO/H_2O as the solvent favoured N-arylation (>50:1) whereas using a weaker base such as K_3PO_4 or Cs_2CO_3 in the presence of ethylene glycol gave a reduced ratio of N/O arylated product and using Cs_2CO_3 in butyronitrile favoured *O*-arylation albeit with lower yields. Steric effects were also observed, using less reactive secondary amino-alcohols increased competition between N and O as the increased steric bulk around the more nucleophilic amino group hinders *N*-arylation.[83] Chain lengths of $n > 3$, between the amino and alcohol groups, give good selectivity of either the amino or alcohol groups depending on the reaction conditions. Chain lengths of $n = 2$ or $n = 3$ give poor selectivity due to the formation of a kinetically favoured 5 or 6-membered ring. As both the N and O atoms are coordinated to Cu, both are susceptible to arylation, thus explaining the poor selectivity.[83]



Scheme 6: N and O selectivity of different ligands for the Ullmann coupling of amino-alcohols with 4-iodo-toluene.

For the application on non-branched amino-alcohols, the use of an additional ligand was required (**Scheme 6**). Using neocuproine (**NEO**) as the ligand in toluene, *O*-arylation is favoured (89-91%, O/N = 18/1-24/1) whereas the use of isopropylcarbonylcyclohexanone (**IPCC**) in DMF favours *N*-arylation (96-99%, N/O = 45/1-50/1).[83] Similar results were also obtained by Chan et al. in 2008 using a ligandless system of solvent and CuI for the arylation of linear amino-alcohols.[84] Changing the solvent showed dramatic changes in reaction selectivity, DMF favoured the formation of the *N*-arylated product, whereas toluene gave only *O*-arylated and *N*,*O*-diarylated product.[84] The role of the ligand on selectivity was explained by the electronic effects of the ligand on the Cu atom (**Figure 1.12**). Buchwald et al. proposed that the anionic ligand (**IPCC**) makes the Cu(I) coordinated species less electrophilic and thus disfavouring the coordination of the hydroxyl group and the more nucleophilic amine group is bound instead. The neutral ligand (**NEO**) makes the Cu(I) coordinated species more electrophilic, favouring coordination of the hydroyxl group over the amine.[83]

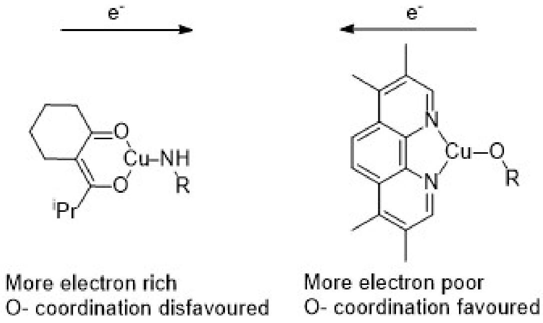
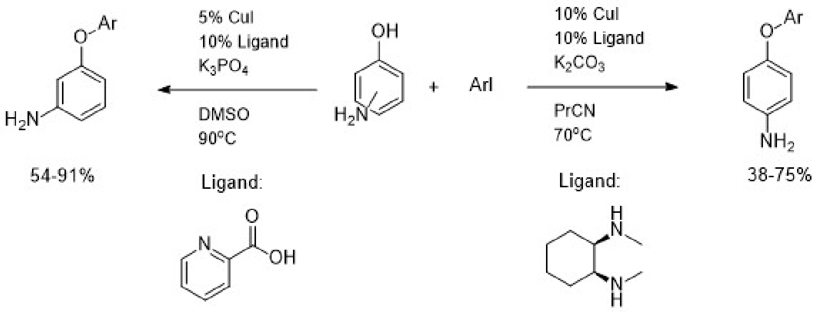


Figure 1.12: Ligand effect on N/O selectivity.

A few years later Buchwald and co-workers published a follow-up computational study to explain the observed selectivity with the above ligands, using methylamine/methanol, aryl iodide and CuI catalyst.[85] The calculations showed that coordination of the O atom was always energetically favoured over the N atom and therefore, the selectivity could not arise during the coordination of the nucleophile to the Cu atom. The selectivity could instead be explained by the activation of the aryl halide. Using **IPCC**, *N*-arylation was always favoured over *O*-arylation, with the lowest energy pathway being an outer sphere single electron transfer mechanism. For neocuproine the *O*-arylation activation energy via an inner sphere iodine atom transfer mechanism was lower than for the *N*-arylation.[85]

A similar computational study was performed by Fu and co-workers on amino-alcohols instead of methylamine/methanol, which is a more realistic system.[86] In their study neocuproine was replaced with the less electron-rich, phenanthroline (**PHEN**). Their calculations suggested that the oxidative-addition/reductive-elimination mechanism was the most favourable, with the selectivity explained by a difference in the order of coordination of the nucleophile and oxidative addition.[86] As **IPCC** is an anionic ligand, the charge on the complex with Cu(I) is 0 and therefore, is more prone to oxidative addition than the complex of Cu(I) with phenanthroline, with a +1 charge. In order for oxidative addition to occur with the Cu(I)−**PHEN** complex, coordination of the nucleophile is required to create a neutral species.[86] As a result, the oxidative addition of the aryl halide to the Cu(I)−**IPCC** complex is relatively easy, with the rate-limiting step being the coordination of the nucleophile to the Cu(III) species, whereas in the Cu(I)−**PHEN** complex the rate-limiting step is the oxidative addition of the aryl halide. As energy is required to coordinate the nucleophile in the Cu(I)−**IPCC** complex the more nucleophilic amino group coordinates selectively, leading to *N*-arylation. The hydroxyl group preferentially coordinates to the Cu(I)−**PHEN** complex due to its higher acidity, leading to *O*-arylation.[86]



Scheme 7: Conditions used by Buchwald and co-workers (2009) for the coupling of aminophenols with aryl iodides.[82]

Buchwald's group expanded their substrate scope to aminophenols in their study in 2009 (**Scheme 7**), which led to the development of conditions which gave almost complete O selectivity in 3-aminophenols.[87] Applying these conditions to 4-aminophenols however, gave much lower yields and selectivities as well as being more sensitive to steric hindrance.[87] Further expansion to 2-aminophenols gave no *O*-arylated product both with and without additional ligands, with only *N*-arylated and *N*,*N*-diarylated products being observed.[87] This suggested that arylation preferentially takes place at the more nucleophilic amino group when both groups are bound to the copper.[87] In a competition study using equimolar amounts of aniline and phenol, however, only phenol-coupled product was observed. When aniline was replaced with less electron-rich aniline derivatives, the *N*-coupled product again became the dominant product.[87] Based on this observation, the authors discussed a mechanism in which deprotonation plays an important role in the catalytic cycle.

The more nucleophilic amino group binds to the Cu atom before deprotonation. Whereas the more acidic phenol is deprotonated at the beginning of the reaction, prior to binding to the Cu atom, and subsequently coordinates in its anionic form faster than the amine due to the presence of the negative charge. Competition between these two complexes is present in the reaction and the selectivity is dependent on their relative rates of formation and oxidative addition. The nucleophilicity and deprotonation rate of the amine shifts the equilibrium between these two species. Electron-poor amines are more competitive due to being more acidic and hence, can be deprotonated before coordination as a negatively charged species.[87]

1.4 Conclusions

Catalysts are a vital component of modern synthetic processes, required in the majority of industrial chemical synthesis procedures. The development of new, highly active catalysts is key to improving industrial processes and enabling the synthesis of previously unavailable chemical compounds. The drive towards sustainability in chemistry has made the need to move from expensive and rare precious metals towards cheaper, readily available base metals. However, the lack of understanding of these catalysts, both mechanistic and ligand understanding, has made the uptake of these metals in catalysis slow.

Structural databases have become an invaluable source of chemical data that is currently underutilised, especially in catalysis. Development of computational workflows that utilise this data will be invaluable for the chemical exploration and prediction of activity for catalysts. Such a process would reduce the number of resources required for route development shortening the time required from lab to consumer. Current approaches to tackling this issue were explored, most of which lack experimental validation or are limited in scope.

Multiple approaches to the computational discovery of novel catalysts are under development. Ligand knowledge bases provide a database of ligands and their properties which can be drawn upon to identify potentially highly active ligands. However, the structures available are limited to those which are currently well-studied or commercially available. Q2MM approaches have only been successfully applied to the prediction of stereoselectivity and have not been applied to the identification of novel ligands. Machine learning methods have started to increase in popularity but their predictions are rarely validated experimentally. Machine learning models can be used to explore the importance of ligand properties and predict specific properties of molecules of organometallic complexes but requires large datasets, which are not readily available, and careful tuning of both the type of model and the descriptor sets used.

The Ullmann-Goldberg reaction is a synthetically important C−N coupling reaction similar to the widely used Pd-based Buchwald-Hartwig reaction, making it a good test reaction to demonstrate the applicability of such an approach. The history of the reaction was explored as well as its advantages and disadvantages compared to palladium. Finally, the current mechanistic and ligand understanding was explored. The prominence of nitrogen and oxygen-based ligands provides a starting point for the targeting identification of novel ligands within structural databases.

1.5 Aims & Objectives

Computational approaches are rarely used in the chemical industry, particularly in the field of catalyst design and selection. Identification and validation of key structural features in organometallic complexes enables the design of catalytic species for numerous applications in silico. A rational approach based on extensive structural/activity data will be essential in reducing resources dedicated to catalyst screening and late-stage development. Higher success rates and reduced development time and resources will enable the catalyst industry to deliver solutions faster and more effectively.

The aim of this work is to apply structure-based design approaches, already established in drug design, to the field of organometallic catalysis. Moving away from toxic and expensive rare-earth metals to safer, cheaper and more abundant base metal catalysts is an attractive prospect for making chemical synthesis more sustainable. However, base metal catalysis is less understood compared to precious metals. A lack of both mechanistic and ligand understanding makes the development of ligands for these catalysts much more challenging. As an example use case the Ullmann-Goldberg reaction is used to identify new ligands to improve the viability of the reaction with respect to the Buchwald-Hartwig reaction preferred in industry.

This will be done in the following steps:

i) Creation of a structural database (**CatSD**) from the Cambridge Structural Database containing features relevant to organometallic catalysts.

ii) Development of a high-throughput computational workflow for the prediction of activation energies. The time scale of which should be similar to experimental high-throughput screening.

iii) Application of **CatSD** and the high-throughput workflow to a reaction of industrial interest and limited ligand and mechanistic understanding. The output of which will be used to link ligand/complex properties to catalytic activity.

iv) The accuracy of the workflow will be assessed via comparison with high-level wave function methods and comparison with experimental data.

v) Machine learning will be used to link ligand steric and electronic properties to activity. Models will also be used to predict activation energies both with and without transition state structures to assess their accuracy and the ability to predict activation energy without the need for calculation of the transition state.

1.6 Thesis Layout

This thesis will follow the semi-automated high-throughput cheminformatic approach to identifying and predicting ligand activity for homogeneous catalysis. **Chapter 2** will describe the process of building the Catalyst Structural Database (**CatSD**), the definition of structural features and using **CatSD** to search the Cambridge Structural Database for potential ligand structures.[88] **Chapter 3** will introduce a new approach to predicting ligand activity through a semi-automated high-throughput computational workflow along with benchmarking data against higher-level wavefunction methods and 3D structural data. **Chapter 4** will describe the application of both the computational workflow and **CatSD** to the Ullmann-Goldberg reaction, where both amine and amide nucleophiles are explored. **Chapter 5** describes machine learning models for the prediction of activation energy, identification of incorrect structures and the importance of ligand properties. Finally, a direction for future work in this field is posited. **Figure 1.13** shows the methodological workflow employed in this thesis and the chapters of the thesis in which they are presented and discussed.

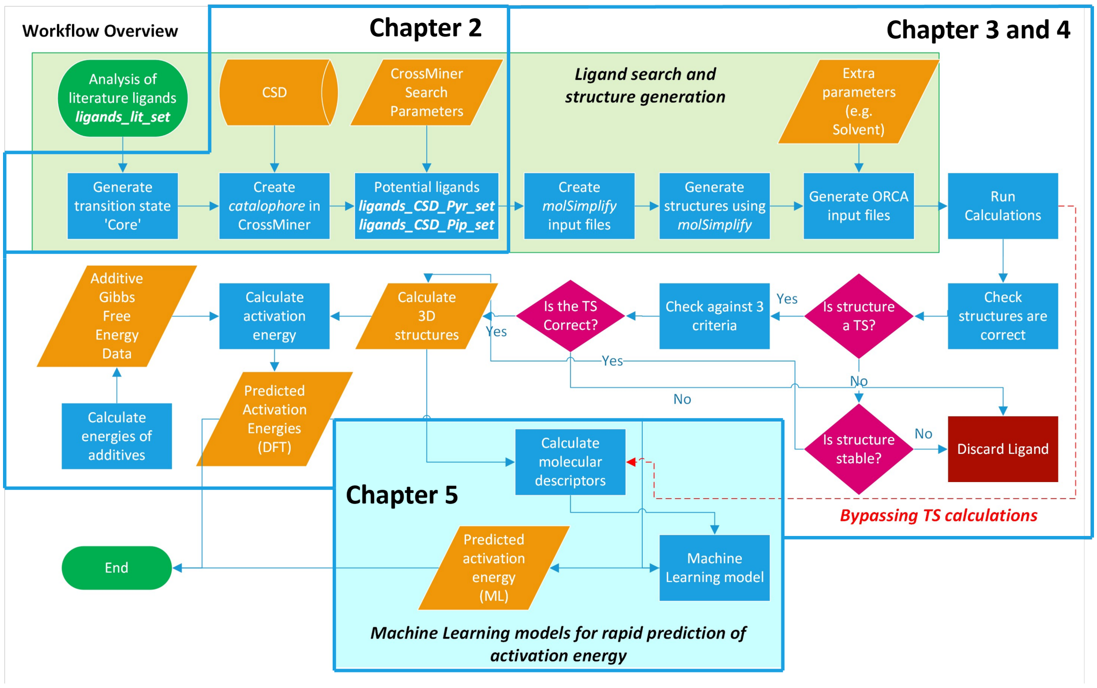

Workflow Overview
Chapter 2
Chapter 3 and 4
Analysis of literature ligands *ligands_lit_set*
CSD
CrossMiner Search Parameters
Ligand search and structure generation
Extra parameters (e.g. Solvent)
Generate transition state 'Core'
Create *catalophore* in CrossMiner
Potential ligands *ligands_CSD_Pyr_set* *ligands_CSD_Pip_set*
Create *molSimplify* input files
Generate structures using *molSimplify*
Generate ORCA input files
Run Calculations
Additive Gibbs Free Energy Data
Calculate activation energy
Calculate 3D structures
Is the TS Correct?
Check against 3 criteria
Yes
Is structure a TS?
Check structures are correct
Yes
Yes
No
No
Calculate energies of additives
Predicted Activation Energies (DFT)
Calculate molecular descriptors
Is structure stable?
No
Discard Ligand
Chapter 5
Bypassing TS calculations
End
Predicted activation energy (ML)
Machine Learning model
Machine Learning models for rapid prediction of activation energy

Figure 1.13: An overview of the new computational and cheminformatic workflow developed in this work and the relevant chapters in which these steps are discussed.

1.7 Data & Code Availability

All curated datasets and code described in this work are freely available at `https://github.com/MarcS18/Thesis_ESI`. The code is split into folders based on the chapter and corresponds to the respective section number within the chapter. The folders also contain sample data to demonstrate how the code works.

Chapter 2: CatSD: Structural Database for Catalyst Design and Development

2.1 Introduction

Crystallographic databases are an invaluable resource for structural-based design due to the breadth of structural data available which covers a large variety of chemical species. Exploitation and interpretation of this large quantity of structural data has led to the generation of practically useful empirical conformation rules and interaction preferences.[89,90] Creation of a structural database built upon the structure-rich CSD would provide a data-rich resource for chemists to perform structure-based design, analysis or predictions for a large variety of chemical problems within catalysis.

2.1.1 Structure-based Design

The CCDC CrossMiner approach is based on searching crystal structure databases such as the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB) in terms of a pharmacophore query (**Figure 2.1**).[91] This approach is primarily used for the structure-based design of pharmaceuticals. Pharmacophore queries aim to describe the protein-ligand interaction patterns, ligand scaffold or protein environments. Example features to describe pharmaceutical properties include hydrogen bond donors/acceptors, planar rings and excluded volumes. The main application of CSD-CrossMiner is as a complementary tool in drug discovery to identify new drug candidates through analysing drug-protein interactions.

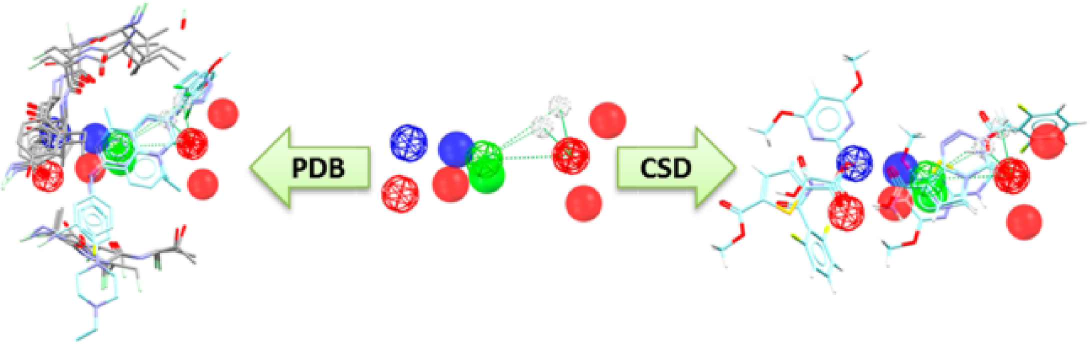


Figure 2.1: Graphical representation of the CrossMiner approach for drug design, where the coloured spheres represent pharmacophore points are used to search the CSD or PDB to retrieve matching chemical structures.[91]

Comparisons can be drawn between drug docking and catalysis. The catalyst can be considered as the protein and the substrate(s), the drug molecule. In this case, the 'inverse' approach is applied. where the substrate(s) (drug) is known and the catalyst (protein) is unknown. The user will define the substrate(s) and the 'catalophore', a set of features used to describe the properties of the ligand(s)/catalyst. Which can be used to search the CSD for potential ligands for the chemical transformation of interest.

Unlike traditional ligand docking, which relies primarily on intermolecular interactions such as hydrogen bonding and ionic interactions, catalysis relies on the electronic and steric environment at the metal centre, the properties of which are influenced by the coordinated ligands. Structural data alone is insufficient to represent the electronic environment and therefore, must be coupled with a computational modelling method such as density functional theory. Density functional theory is a time-consuming method so a faster method is required to screen large numbers of molecules in a much shorter time frame in order for it to be considered as an alternative to experimental screening approaches.

CSD-CrossMiner for catalysis would operate from a transition state where the user inserts their substrate on a transition state template (reference structure), upon which the CrossMiner pharmacophore features can be applied to the ligand(s) and the database searched to retrieve similar compounds which could be used as potential ligands. The generated structures will then either be exported and used for either DFT calculations or input directly into a machine learning model to directly predict the activity of the ligand(s)/catalyst.

2.1.2 CSD-CrossMiner Terminology

2.1.2.1 Feature and Pharmacophore points

Features are a point, centroid or vector placed in 3D space which represents a SMARTS query to match when searching for matching structures. If the feature is a vector it also includes geometric rules to define the directionality of the feature. A single-point feature is represented as a single small translucent sphere. A directional feature is represented by two types of spheres, *base* and *virtual*, which are displayed as small translucent spheres. The *base* feature represents the feature itself, while the *virtual* point(s) represent the directionality of the feature. Directional features can have one or more virtual spheres to represent different geometries (1 for a linear feature, 3 for a tetrahedral feature etc.).

A pharmacophore point is a feature that is necessary to ensure optimal interactions or structure with a specific target. For example, in medicinal chemistry, the pharmacophore point is a feature that is necessary for optimal intermolecular interactions with a specific biological target to trigger or block its biological response. A pharmacophore point is represented as a mesh sphere. The sphere radius of each pharmacophore point represents the tolerance radius and reflects the uncertainty in the position of the pharmacophore point. The radius of a pharmacophore point can be varied to control the specificity of the query. Like with features a directional pharmacophore point can also be used to represent the directionality of a pharmacophore point using *virtual* sphere(s). In this chapter, this will be referred to as a catalophore point.

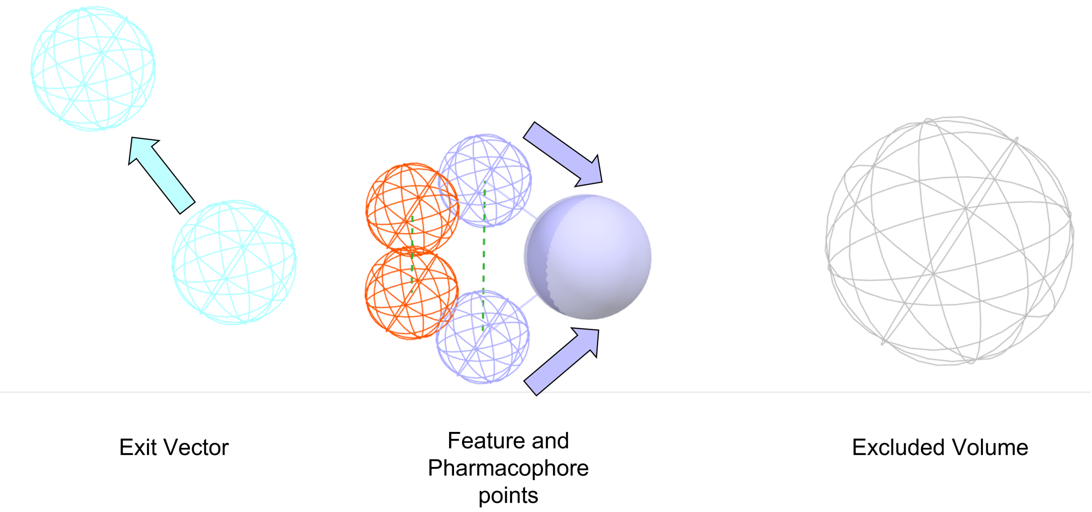


Figure 2.2: Visualisation of feature (orange) and pharmacophore points (purple), with directionality (meshed=base, solid=virtual), excluded volumes (white) and exit vectors (light blue) in CSD-CrossMiner. The green dashed line represents that features are constrained intramolecularly.

2.1.2.2 Excluded Volumes and Exit Vectors

An excluded volume is a feature that defines a volume in 3D space whereby no atom can be present. An excluded volume is represented by a single mesh sphere. In medicinal chemistry, an excluded volume feature defines an occupational volume where no solute molecule can be present.

An exit vector is a two-point feature that represents a single, non-ring bond between two heavy atoms, and is used to represent an R group of any composition or a branching point in a structure. In CSD-CrossMiner an exit vector is bi-directional and therefore has no directionality. An exit vector is represented as two mesh spheres.

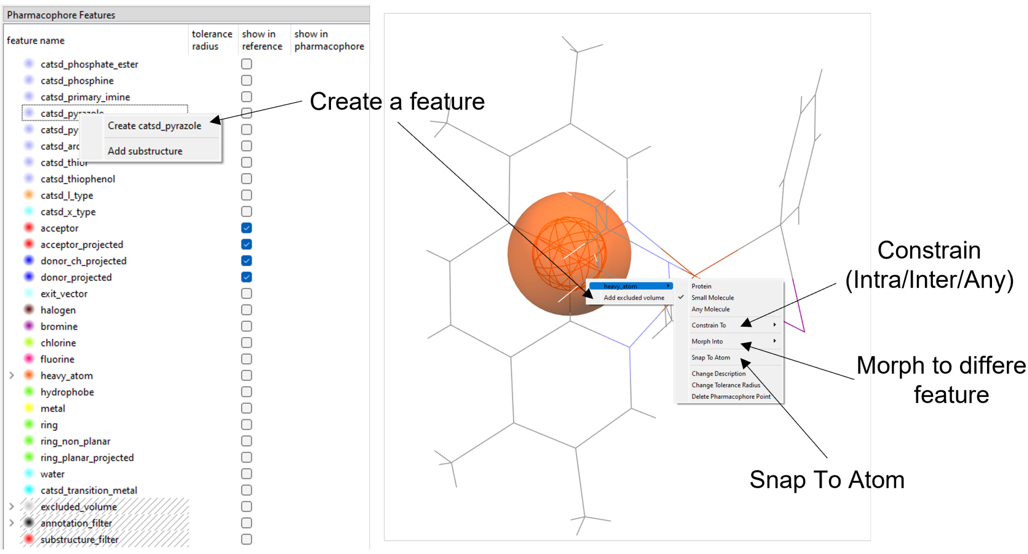


Figure 2.3: Feature selection pane as seen in CSD-CrossMiner.

Search terms are generated by adding features from the pharmacophore features pane in CSD-CrossMiner. Features can be added either by right-clicking a matching atom (displayed by a small sphere) on a reference structure or by right-clicking the desired feature on the feature pane (**Figure 2.3**) and moving the sphere onto the atom of interest. Feature spheres can be snapped to atoms by right-clicking the sphere and selecting ‘snap to atom’, which will move the sphere to the nearest atom. Features can also be constrained to be either intermolecular/intramolecular or to any other feature in the catalophore. Intramolecular constraints ensure that all features belong to the same structure in the hit compounds.

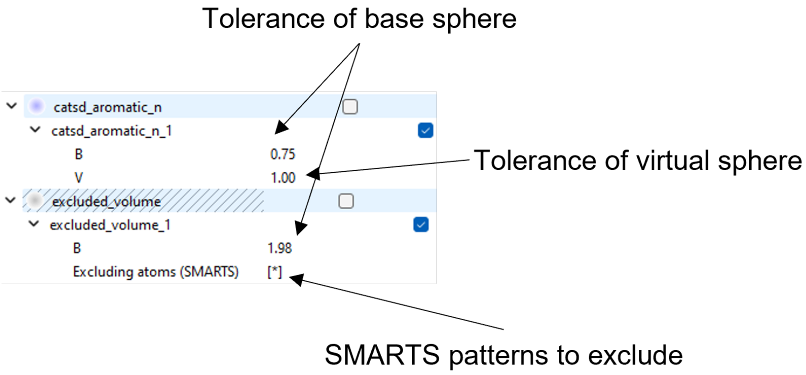


Figure 2.4: Expanded view of a feature in the feature pane.

The tolerances of a sphere can be changed by expanding the feature in the feature pane and altering the values, B for the *base* sphere and V for the *virtual* sphere (**Figure 2.4**). For excluded volumes specific SMARTS patterns can also be excluded (* excludes all atoms).

2.1.2.3 Structure and Feature Databases

A structural database contains the 3D coordinates of a set of chemical structures. Structural databases can be generated straight from the CSD or from other community or proprietary datasets. Structural databases are used as a basis to generate a feature database.

A feature database contains the 3D structures from the structural database but is indexed with a set of feature definitions. Indexing matches the features in the feature database to each structure in the structural database allowing them to be searched using the feature points in a search query. Either the default feature definitions can be used or custom definitions input by the user. This is the database that CSD-CrossMiner uses to perform the 3D search against the query.

2.2 Results and Discussion

2.2.1 Building CatSD

The structural database is built upon CSD_541 with the Mar20, May20, Aug20 and Feb21 updates. **CatSD** contains structures from the CSD which (a) are not polymeric, (b) have no disorder, (c) for which 3D coordinates have been determined and (d) have a maximum *R*-factor of 10%. This was to ensure that all structures present in the database had high-quality 3D coordinates and to remove any polymeric structures, which are not relevant to homogeneous catalysis. This resulted in a database containing approximately 658,000 structures.

To enable easier searching of the database a set of annotations are used to define specific properties of each entry in the database. Each entry has a set of annotations taken directly from the CSD, using the CSD Python API, containing the following values: (1) CSD identifier, (2) CSD Refcode, (3) database name, (4) chemical formula, (5) *R*-factor, (6) is_organic and (7) is_organometallic. These annotations allow the database to be searched in terms of the quality of the 3D structure (*R*-factor) as well as whether the structure is organic or organometallic. The ability to search for only organic structure is very useful due to several issues with searching for localised features with CSD-CrossMiner (see **Section 2.2.2**).

CatSD itself is built from the Cambridge Structural Database however, the features included within the database can be used for custom databases; both freely available and proprietary through CrossMiner using the database generation tool. Custom annotation sets can also be used by adding a custom *.csv* file containing the required annotations for each structure. The script for extracting annotations from the CSD using the CSD-Python API is available in the ESI.

2.2.2 Structural Features for Catalysis

Each entry in the structural database is annotated with a set of feature points which are used to perform a 3D search. Each feature is defined by a hierarchy of SMARTS patterns which must be tailored for each use case in order to represent different functionality within a molecular structure.

The default feature set packaged with CSD-CrossMiner is insufficient for catalysis as it does not contain any features which are related to common metal-coordinating functional groups. Several features included with CSD-CrossMiner are useful for catalysis such as *ring* and *ring_projected* features for describing both aliphatic and aromatic ring systems, *heavy_atom* for defining the location of heavy atoms (not H) and *donor_projected* and *acceptor_projected* for describing hydrogen bonding, either intra-ligand or between ligand and substrate.

A new set of features were created for catalysis to enable searching of the CSD for common coordinating functional groups used in organometallic chemistry. Each functional group was defined by a SMARTS string then the SMARTS list expanded whenever an assignment was considered incorrect when tested on a subset of CSD structures from the CSD Aug20 update (**Figure 2.6**). **Table 2.1** shows the full list of SMARTS strings used in the *catsd_-coordinating_atom_general* feature.

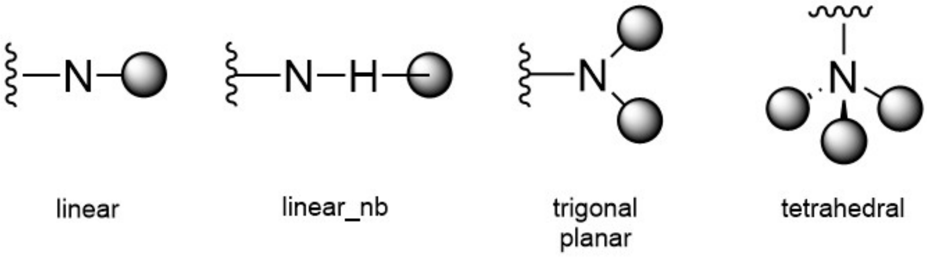
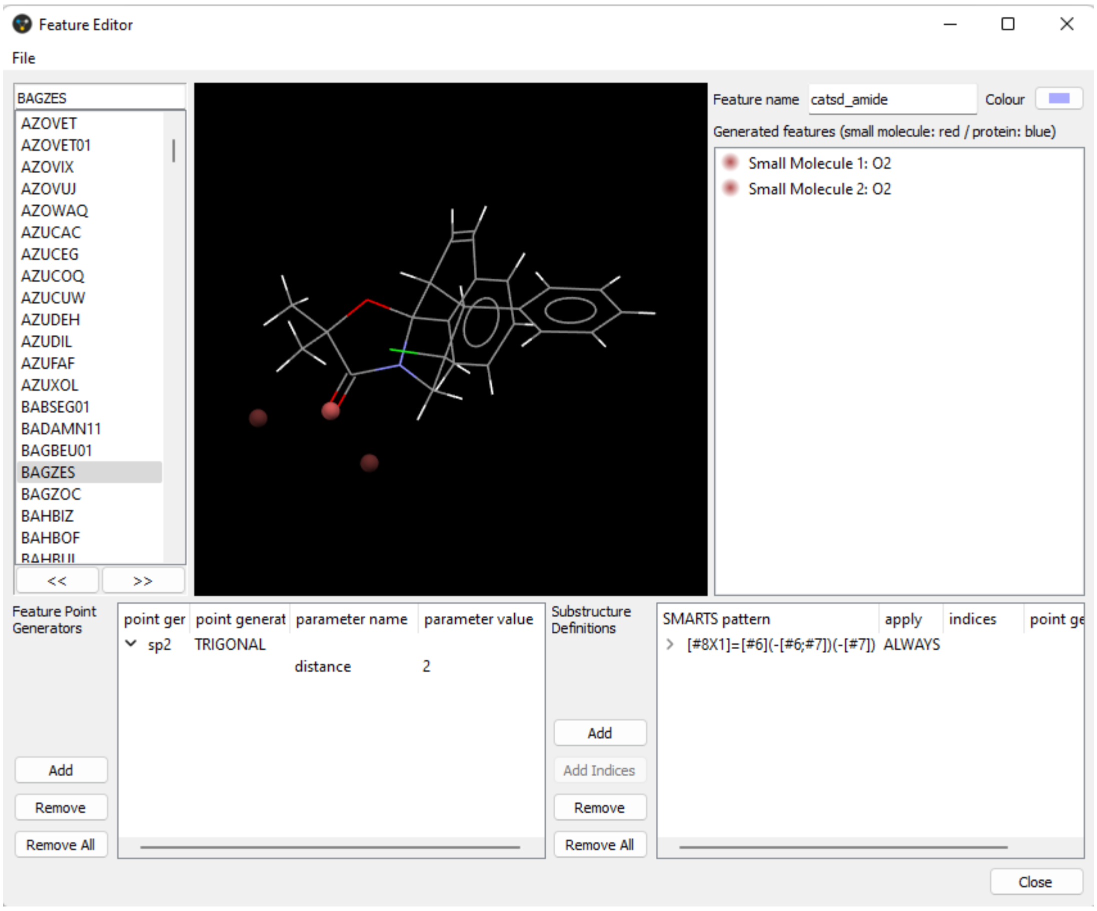


Figure 2.5: Geometries of the features used in CSD-CrossMiner.

Every feature in **CatSD** is a directional feature to ensure that the coordinating functional groups matched are in the correct orientation to allow for binding to the metal centre. All possible geometries and projections are shown in **Figure 2.5**. The geometry of the feature is not directly related to the geometry at the base atom. This is due to the way that CSD-CrossMiner calculates the projected points. For example, for a primary amide (N) a trigonal feature would have the projected points located on both nitrogen hydrogen atoms. However, as both hydrogens are occupying those positions the trigonal feature cannot be used.



Feature name catsd_amide

Generated features (small molecule: red / protein: blue)

- Small Molecule 1: O2
- Small Molecule 2: O2

point ger	point generat	parameter name	parameter value
sp2	TRIGONAL		
		distance	2

SMARTS pattern	apply	indices	point ge
[#8X1]=[#6](-[#6;#7])(-[#7])	ALWAYS		

Figure 2.6: Example feature definition with the directionality of an amide functional group displayed on a matching structure in the CSD.

It is common for structures in the CSD to contain functional groups that have been protonated in the crystal structure. In these cases, the SMARTS string also needs to match the hydrogen atom. An example of this is pyridine where the SMARTS pattern for the nitrogen atom is, n, for the nitrogen or, n-[#1], if the nitrogen is protonated. In this case, a *linear_nb* geometry is used to project the feature from the hydrogen instead. If a *linear* feature is used the hydrogen atom will occupy the projected position and the structure would not be matched by any search.

Table 2.1: Features used in the CatSD *coordinating_atom_general* feature.

Number	SMARTS Definition	Base Index	Geometry	Functional Group
1	[#6](-[#6])(-[#6]=[#6](-[#6])-[#8-])=[#8]	5, 6	trigonal	2,5-diketone
2	[#8-]-[#6]	0	tetrahedral	alcohol
3	[#8H1]-[#6X4]	0	tetrahedral	alcohol
4	[#1][#8H1]-[#6X4]	0	linear_nb	alcohol
5	[#8X1]=[#6](-[#6;#7])(-[#7])	0	trigonal	amide
6	[#7X2]-[#6](=[#8])	0	trigonal	cyclic amide
7	[#1][#7X3H2]-[#6](=[#8])	0	tetrahedral	primary amide
8	[NX3H2](-[C]([!#8;!#7])([!#8;!#7]))	0	tetrahedral	primary amide
9	[#1][#7X3H1]-[#6](=[#8])	0	linear_nb	secondary amide
10	[NX3H2](-[C]([!#8;!#7])([!#8;!#7]))	0	tetrahedral	secondary amide
11	[NX3H0](-[C]([!#8;!#7])([!#8;!#7]))(-[C]([!#8;!#7])([!#8;!#7]))-[C]([!#8;!#7])([!#8;!#7])	0	tetrahedral	tertiary amide
12	[#1][NX3H2](-[c]([!#8;!#7])([!#8;!#7]))	0	linear_nb	aniline
13	[#1][NX3H2](-[c]([!#8])([!#8]))	0	linear_nb	aniline
14	n-[#1]	0	linear_nb	aromatic nitrogen
15	n	0	trigonal	aromatic nitrogen
16	[#7]=[#7][#6]	1	trigonal	azo
17	[#7]=[#7][#6]	0	tetrahedral	azo
18	[#8H1][#6]=[#8]	0	trigonal	carboxylic acid
19	[#1][#8H1][#6]=[#8]	0	linear_nb	carboxylic acid
20	[#1][#8H1][#6]=[#8]	3	trigonal	carboxylic acid
21	[#8X1][#6]=[#8]	2	trigonal	carboxylic acid

Continuation of Table 2.1

Number	SMARTS Definition	Base Index	Geometry	Functional Group
22	[#8X1][#6]=[#8]	0	tetrahedral	carboxylic acid
23	[#6](-[#6])(-[#8]-[#6])=[#8]	4	trigonal	ester
24	[#7]=[#6]-[#7][#1]	3	linear_nb	imidazole
25	[#6;#1;#8;#7;#16][#7X2]=[#6]	1	trigonal	imine
26	[#6]=[#7X3](-[#6;#1;#8;#7;#16])-[#1]	3	linear_nb	imine
27	[#6]=[#7]-[#6]	1	trigonal	imine
28	[#6]#[#7]-[#6]	0	linear	isocyano
29	[#8X1]=[#6]([#6])([#6])	0	trigonal	ketone
30	[#8-]-[#7]-[!#8]	0	trigonal	n-oxide
31	[#1]-[#8]-[#7]=[#6]	0	linear_nb	oxime
32	[#8-]-[#7]=[#6]	0	tetrahedral	oxime
33	[#1]-[#8]-[#7]=[#6]	1	tetrahedral	oxime
34	HOc	1	tetrahedral	phenol
35	HOc	0	linear_nb	phenol
36	[PX3](-[#8H0][#6])(-[#8H0][#6])-[#8H0][#6]	0	tetrahedral	phosphate ester
37	[PX3](-[#6;#7])(-[#6;#7])-[#6;#7]	0	tetrahedral	phosphine
38	[#7X2H1]=[#6]	0	tetrahedral	primary imine
39	[#7][#7][#1]	2	linear_nb	pyrazole
40	c-[#7](-[#1])-(c)	2	linear_nb	pyrrole
41	c[nX2]c	1	trigonal	aromatic nitrogen ring
42	[#16-]-[#6]	0	tetrahedral	thiol
43	[#16H1]-[#6X4]	0	tetrahedral	thiol
44	[#1][#16H1]-[#6X4]	0	linear_nb	thiol
45	HSc	1	tetrahedral	thiophenol
46	HSc	0	linear_nb	thiophenol
47	[NX3H2]([#6]([!#8;!#7])([!#8;!#7]))	0	tetrahedral	primary amine
48	[NX3H1]([#6]([!#8;!#7])([!#8;!#7]))([#6]([!#8;!#7])([!#8;!#7]))	0	tetrahedral	secondary amine

Continuation of Table 2.1

Number	SMARTS Definition	Base Index	Geometry	Functional Group
49	[NX3H0](-[#6]([!#8;!#7])([!#8;!#7]))(-[#6]([!#8;!#7])([!#8;!#7]))-[#6]([!#8;!#7])([!#8;!#7])	0	tetrahedral	tertiary amine

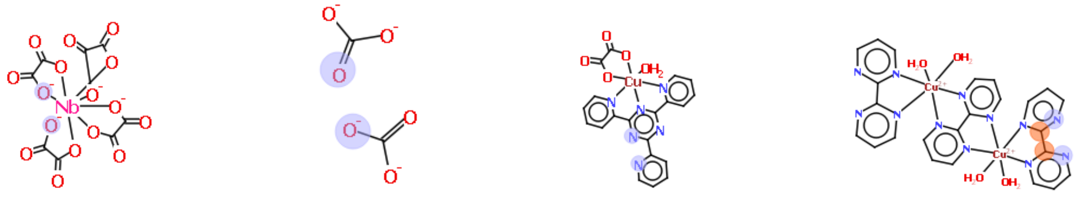
The general coordinating atom feature is a very general feature that broadly searches all coordinating atom types. For a more specific search query each functional group is divided into separate individual features (e.g. primary amide and phosphine). All features included with **CatSD** are indexed and stored in the feature database so are ready to use upon loading the database.

2.2.3 Using CatSD with CSD-CrossMiner

Approximately 57% of the structures contained in the CSD are metal-organic based.[92] Utilising this data however is much more challenging as there are several issues with both how the structures are stored and how the CSD software accesses and performs search queries.

Firstly, CSD-ConQuest, a 2D-based structure searching software for the CSD, contains a search term, ‘4M’, which is a general term for any transition metal. This term is not available in CSD-CrossMiner, nor does SMARTS contain any similar value for defining transition metal elements. To match the SMARTS string for a coordinating atom in an organometallic structure it must include the metal atom as well. This makes the generation of features for coordinating atoms very difficult. Each feature in a catalophore point either needs to contain a list of every single transition metal element per feature or needs to have a separate feature for each transition metal. This is impractical and very time-consuming to generate. In some cases you may also wish to search for structures with a specific transition metal element only, meaning new features will have to be generated and indexed into a new database each time, meaning the features are not transferable across metals. A substructure filter could also be used but this would involve having to list every unwanted transition metal element as an exclusion rule, which is not practical from an ease-of-use standpoint.

Secondly, organometallic structures are considered as a single structure in CSD-CrossMiner. Therefore, confining features to be intramolecular does not confine the features to within a single ligand. This results in a large quantity of hit structures, especially for polydentate ligands, where the two catalophore points are located on different ligands (**Figure 2.7a, 2.7b**). These hits are not viable, as they do not match the desired denticity due to each coordinating atom being located on separate molecules.



(a) Atoms in different ligands

(b) Atoms in different molecules

(c) Incorrect binding mode

(d) Multiple metal atoms

Figure 2.7: Examples of incorrect hits when searching organometallic structures using CatSD.

Thirdly organometallic structures require a lot of pre-treatment before they can be used for activity prediction. Active catalytic states are hard to obtain crystal structures for, therefore, the structures present in the CSD are either pre-catalysts or stable complexes of the metal of interest. This complex must then be converted into the active catalytic form before it can be used for activity prediction. This involves the automated removal of any unwanted ligands, which is a difficult and time-consuming task.

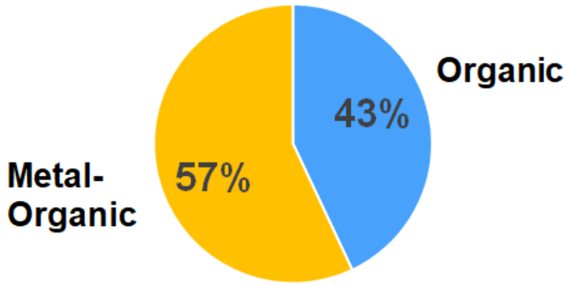


Figure 2.8: Percentage of structures in the CSD that are organic and metal-organic.[92]

A much more attractive approach is to use the other 43% of the CSD containing organic compounds and search for free ligands instead. For free ligands, features do not require any transition metals to be included and one feature or set of features can be used to describe one functional group across all transition metals. Removing the need for re-indexing every time a new metal is studied and the generation of complex feature SMART strings. As organic compounds only contain the structure of the free ligand, intramolecular confinement of features now works on the ligand only and the hits returned contain just the ligand structure of interest. Finally, as there are no other components in the structure no further treatment is required and the ligand structure can be used directly for ligand activity prediction.

This approach is limited to ligands that can be crystallised as free ligands and not those that can only be crystallised within an organometallic complex. Furthermore, free ligands possess a higher degree of conformational flexibility and may exist in the CSD in a conformation that is not the same as the one when bound to a metal centre. As CSD-CrossMiner uses a 3D space search any ligand in an unfavourable binding conformation will not be matched by the catalophore. Therefore, some ligands that are only accessible in organometallic structures will not be identified when searching the CSD.

One possible approach to circumvent this issue is to create a custom database, whereby all of the structures are from a DFT optimized structure where every ligand is bound to simple metal complex such as $PdCl_2$ and remove the $PdCl_2$ moiety after. This will provide a database with every free ligand in to correct binding conformation. However, this will be extremely expensive both with computational resources and time. Such databases could also include the curation of commercially available ligands to facilitate high-throughput prediction in process chemistry and route selection.

2.2.3.1 Generating Catalophores

In order to search the CSD using CSD-CrossMiner a query needs to be generated describing the key 3D structural features the structure must possess. The default name for the query in CSD-CrossMiner is a pharmacophore as it is used primarily for the identification of potential drug candidates. Due to the differences between pharmaceuticals and catalysts and to discern between the two, we use the term ‘catalophore’ to describe the search query. An overview of the process to generate a catalophore is shown in **Figure 2.9**.

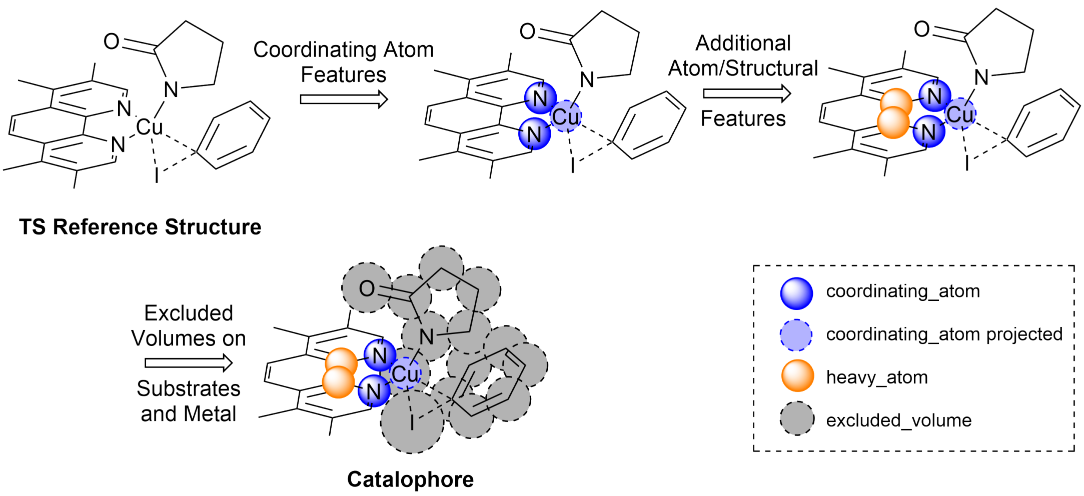

TS Reference Structure
Coordinating Atom Features
Additional Atom/Structural Features
Excluded Volumes on Substrates and Metal
Catalophore
O
N
Cu
I
coordinating_atom
coordinating_atom projected
heavy_atom
excluded_volume

Figure 2.9: General workflow for the generation of a catalophore from a transition state reference structure.

In order the generate the catalophore search query a reference structure should be used to provide a basis for placing catalophore points. Example reference structures include an active catalytic state or transition state. **Figure 2.10** shows an example reference structure for the oxidative addition transition state for the Ullmann-Goldberg reaction with a phenanthroline-based ligand.

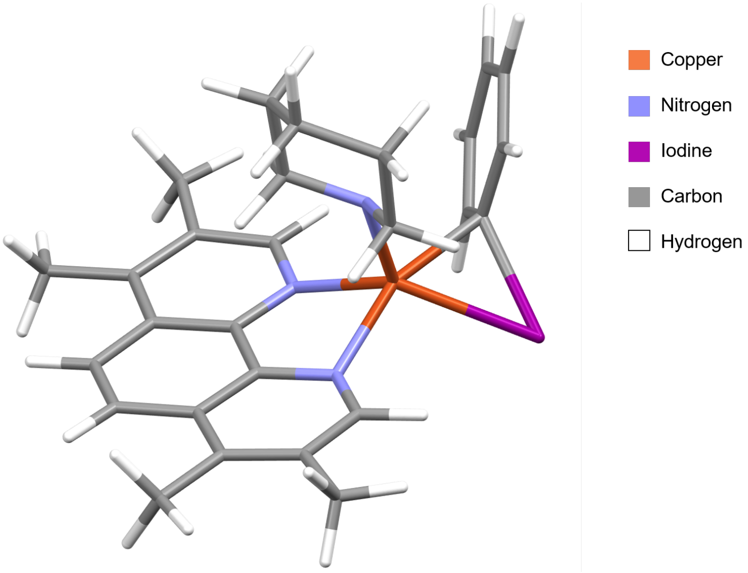


Figure 2.10: Example transition state reference structure.

To define the coordinating atoms of a ligand the `catsd_*` features, included in the feature database, should be used. The catalophore meshed *base* sphere should be placed on the coordinating atom in the reference structure and the solid *virtual* sphere should be placed on top of the metal centre. This determines the directionality of the coordinating atom, ensuring that it is in the correct conformation to enable binding to the metal. The tolerance on the spheres can be adjusted to change the strictness of the search. A smaller tolerance (Å) on the sphere enforces a smaller deviation between catalophore and hit and therefore, a stricter coordination geometry. An example coordination environment in a catalophore is shown in **Figure 2.11** for a bidentate ligand using two `catsd_coord_atom_general` features projecting onto a copper centre.

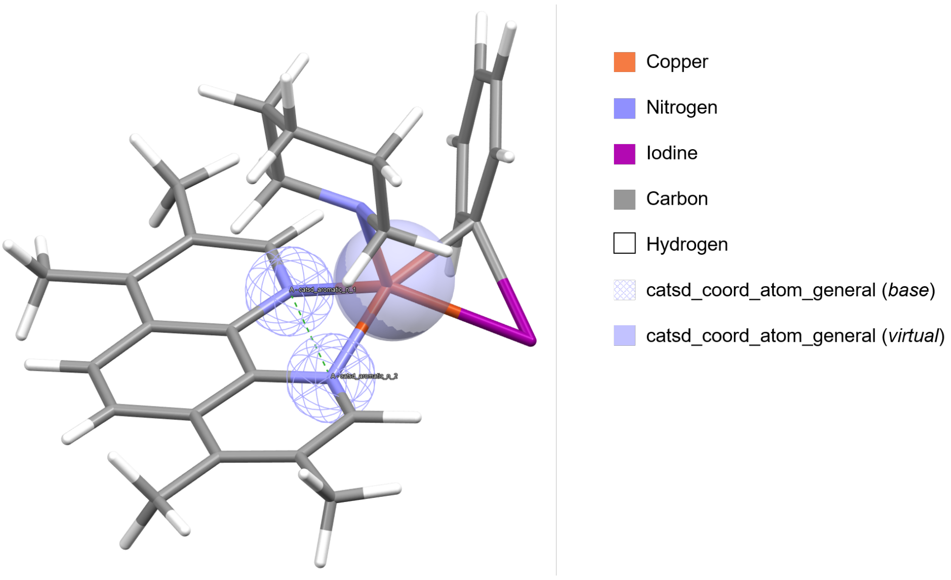

🟧 Copper
🟦 Nitrogen
🟪 Iodine
■ Carbon
□ Hydrogen
▨ catsd_coord_atom_general (*base*)
▧ catsd_coord_atom_general (*virtual*)

Figure 2.11: Catalophore with only coordinating atom features.

Additional structural features for the ligand can be added using CSD-CrossMiner's standard features such as heavy atoms, hydrogen bonds and planar/non-planar ring systems. In **Figure 2.12** a two-atom bridge has been defined using `heavy_atom` features (orange) to create a 5-membered coordination ring. Three `ring_planar_projected` features are used to identify similar structures in the CSD with a phenanthroline ring system motif. Features should be connected intramolecularly to avoid hit structures where the coordinating atoms or other features are present in two different components.

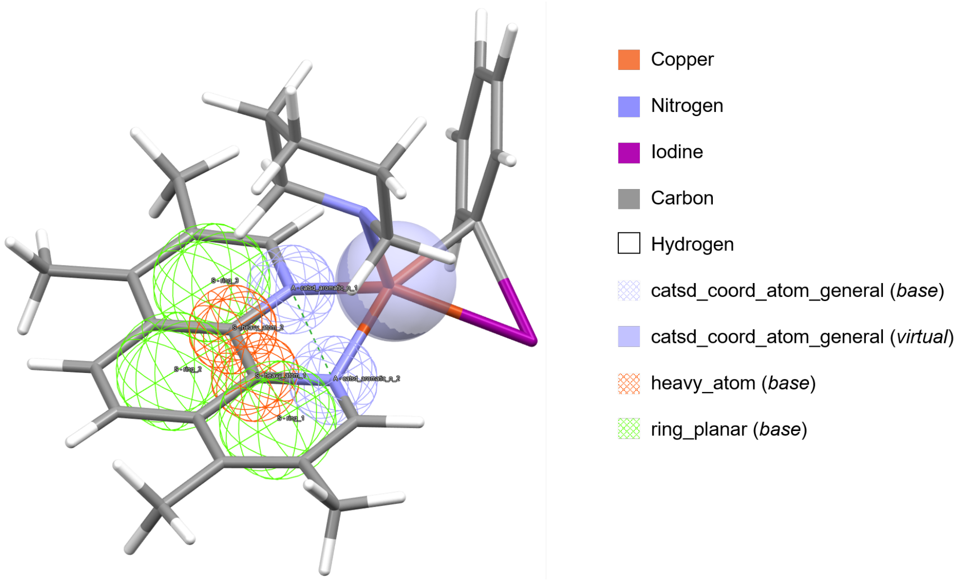


Figure 2.12: Catalophore with coordinating atom and ligand features.

When identifying potential ligands, ideally, the ligands should not occupy the positions taken by the substrates due to high steric crowding, to allow enough space around the metal centre for the substrates to bind. To define the substrate sites excluded volumes are used. Excluded volumes should be placed on the substrates in such a manner that defines an area of space that the ligand should not occupy. Excluded volumes are soft tolerances and therefore, the van der Waals radii of an atom may enter the volume occupied by an excluded volume feature. This can be prevented by increasing the tolerance on the excluded volume features. **Figure 2.13** shows an example catalophore with excluded volumes added on all substrate atoms, with a tolerance of the van der Waals radii of the base atom.

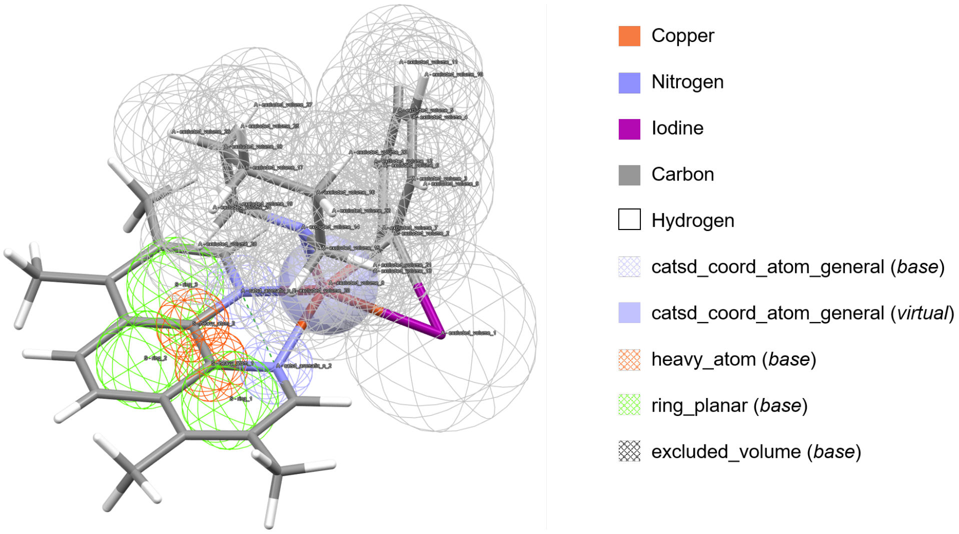
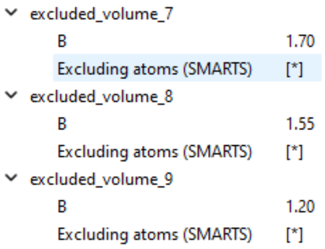

🟧 Copper
🟦 Nitrogen
🟪 Iodine
⬛ Carbon
□ Hydrogen
▨ catsd_coord_atom_general (*base*)
▨ catsd_coord_atom_general (*virtual*)
▨ heavy_atom (*base*)
▨ ring_planar (*base*)
▨ excluded_volume (*base*)

Figure 2.13: Catalophore with ligand features and substrate excluded volumes.

Excluded volume features should have a tolerance of the van der Waals radii of the base atom (e.g. 1.55Å for N) or greater. SMARTS strings can also be used to exclude specific functional groups from the substrate sites in niche cases, instead of all atoms.



excluded_volume_7

B	1.70
Excluding atoms (SMARTS)	[*]

excluded_volume_8

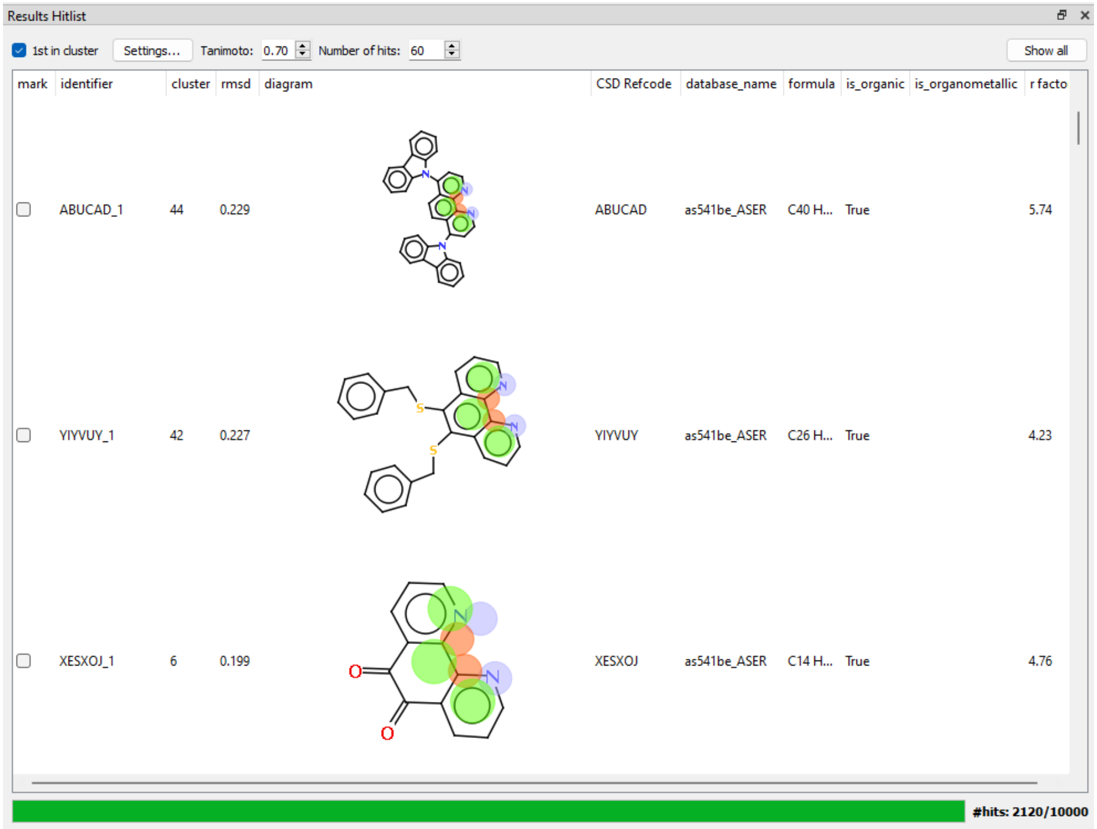
B	1.55
Excluding atoms (SMARTS)	[*]

excluded_volume_9

B	1.20
Excluding atoms (SMARTS)	[*]

Figure 2.14: Variables used when defining an excluded volume in CSD-CrossMiner.

A complete catalophore can then be used to search the CSD for structures matching the search query. **Figure 2.15** shows an example hitlist from the above catalophore. 2120 hits were identified from within the CSD with the same structural motif. Example hits show several results with the same phenanthroline ring system motif, along with a two-atom bridge and two coordinating atoms in the correct orientation to allow binding to a metal centre.



Results Hitlist

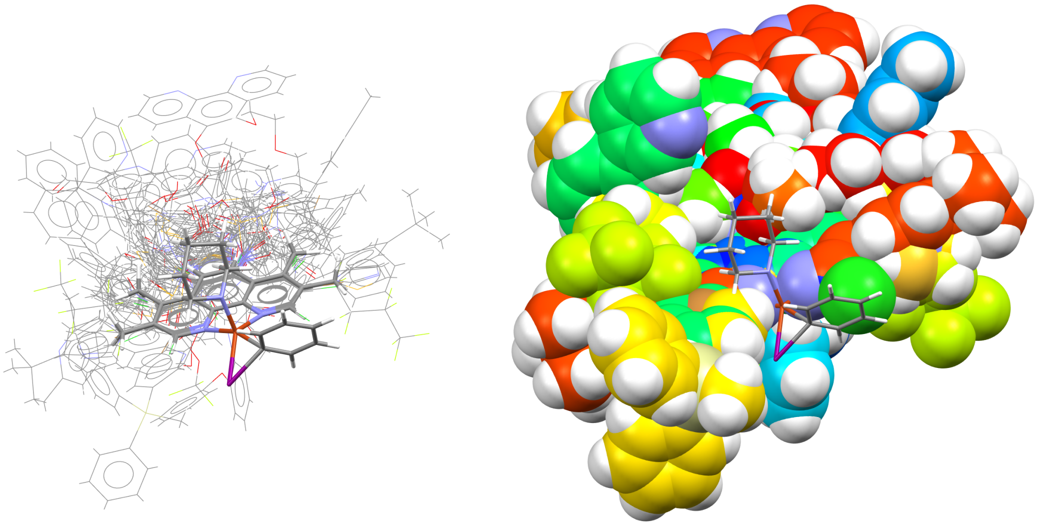
☑ 1st in cluster | Settings... | Tanimoto: 0.70 | Number of hits: 60 | Show all

mark	identifier	cluster	rmsd	diagram	CSD Refcode	database_name	formula	is_organic	is_organometallic	r facto
□	ABUCAD_1	44	0.229		ABUCAD	as541be_ASER	C40 H...	True		5.74
□	YIYVUY_1	42	0.227		YIYVUY	as541be_ASER	C26 H...	True		4.23
□	XESXOJ_1	6	0.199		XESXOJ	as541be_ASER	C14 H...	True		4.76

#hits: 2120/10000

Figure 2.15: Example hitlist from CSD-CrossMiner.

When looking at the structures in 3D space, the majority of structures possess the phenanthroline ring system motif (**Figure 2.16a**) with varying R groups branching off of the ring system. Using the space fill view (**Figure 2.16b**), to view the space occupied by the ligands around the metal, we can see that no ligands are protruding into or occupying the substrates sites. Hits can be exported as a *.csv* file or as 3D structures and visualised in external software or in Mercury for additional analysis.



(a) Wireframe **(b)** Space fill - coloured by hit cluster

Figure 2.16: CrossMiner hits overlaid on the reference structure. Left: Wireframe showing the structural similarity between reference and hits. Right: Spacefill showing the open substrate sites, ligands are coloured by hit cluster.

While the CSD-CrossMiner interface is a useful tool for creating and visualising the search results, the data that can be exported using CSD-CrossMiner's default interface is insufficient for use in building structures for pathway exploration or property prediction as extra information is required. In order to perform the search and retrieve the extra information regarding coordinating atoms the CSD-PythonAPI is used.

2.2.3.2 Searching the CSD via the Python API

Once a catalophore has been generated using the CrossMiner GUI it can be used to search the CSD via **CatSD** to identify potential ligands. Searching must be conducted via the CSD-Python API due to the requirement to extract additional information about the coordinating atoms which is unavailable via the GUI.[93] CSD-CrossMiner only return the structure of the hit. Solvents, salts and any other structures in the crystal structure are not returned and therefore minimal treatment is necessary. An example script is provided in the supplementary information available on GitHub (**Appendix 2.A**). **Note:** A CSD-Discovery license is required to use both CSD-CrossMiner and the following search script.

2.2.3.2.1 Search Settings

The script supports the following arguments to adjust the search procedure:

```
-n, --name
```

The name of the search, which is applied as a prefix to all output files. Value=str.

```
-d, --database
```

The feature database to search. The **CatSD** feature database should be used. The value should be a file path to the location of the feature database, e.g. './CatSD.feat' if the feature database is in the same folder as the search script. If no feature database is supplied the script will default to using the standard CSD-CrossMiner feature database.

```
-c, --catalophore
```

The catalophore file, e.g. 'example_catalophore.cm'. Values include the text string of the name of the catalophore file or the file path of the catalophore file.

```
-m, --max-hit-structures
```

The maximum number of results to return from a search. Default=50000. Value=int.

```
-r, --rmsd
```

The maximum value of the rmsd between the catalophore and the hit structures. Default=1. Value=float.

```
-w, --max-molecular-weight
```

The maximum molecular weight of the hit structures. Default=500. Value=int.

```
-t, --threads
```

The number of CPU threads to use for the search. Default=4. Value=int. An example input is shown below:

```
python cm_search.py -n "example_search" -c "example_catalophore.cm"
-d "./CatSD.feat" -t 8 -m 10000
```

This will search the **CatSD** feature database using the 'example_catalophore' catalophore with a maximum number of 10,000 hit structures using 8 threads and providing output files with the prefix 'example_search'. Further search refinement can be used to alter the search procedure by modification of the python script. The following settings are used by default:

```
    searcher.settings.max_hits_per_structure = 1

    searcher.settings.three_cubed_packing = True

    searcher.settings.complete_small_molecules = True
```

By default, the script only returns a maximum of one hit per structure. This is to reduce the number of duplicate structures returned by the search. For some use cases, such as comparing coordinating sites within the same ligand, this may not be desirable and should be either increased or removed. Three cubed packing (3x3x3 packing) is enabled, which restricts the search to 26 unit cells around the central unit cell. This allows for symmetry-related copies of the feature points to be considered for a small molecule crystal structure that matches. Complete small molecules is also enabled, this ensures that the entire molecule is returned and not just the section that is within the catalophore bounding sphere. In most cases, this should be kept enabled so that all the structures returned are complete and can be used directly for structure building. Additional search settings can be found in the CSD-PythonAPI documentation.[93]

2.2.3.2.2 Annotation Filters

Hit structures can be filtered based on the annotations present in the feature database in two ways. First, the annotation can be defined in the catalophore file (see **Section 2.2.3.1**). Second, the annotation filter can be applied within the search script. Annotation filtering is a textual filtering rule that must be defined within the search query, using values that are present within the feature database. An annotation filter consists of a *'key'* and a *'value'*, where the *key* corresponds to the annotation name and the *value* corresponds to the value for each structure for that annotation. A mismatch between the annotation filter *value* and the *value* for a specific entry will result in the hit not being returned. Below is an example of how to apply an annotation filter within the script.

```
model.add_feature(Pharmacophore.AnnotationFilter("is_organic", "True"))
```

This will filter the hits from the search to ensure that all the structures match the `is_organic=True` annotation. This will return only organic structures from within the database. Custom annotations can be used with custom databases to enable personalised filtering of hits.

2.2.3.2.3 Structure Filters

Hits can also be filtered based on chemical structure. Structural filtering is useful in cases where the presence of a certain element in a ligand could potentially lead to prominent side reactions. For example in the Buchwald-Hartwig cross-coupling reaction, the aryl halide coupling partner contains an I, Br or Cl atom, which is involved in the oxidative addition step.[94,95] If any I, Br or Cl atoms are present in any returned ligands this could lead to significant side reactions between the nucleophilic coupling partner and the ligand or oxidative addition of the ligand onto the Pd centre. Therefore, the exclusion of these structures is highly recommended.

A considerable of structures present in the CSD contain Group 1 and Group 2 metal ions within their crystal structures. This is due either to the conditions used to create the crystal or that they have been included intentionally. Such structures are not suitable to be used as ligands without the time-consuming treatment of the 3D structures. Unwanted elements need to be removed and indexes of atoms adjusted to account for the removal of these atoms. If the structure contains more than one ligand (e.g. $[NaL_3]^+$, where L is a bidentate ligand) the additional ligands also need to be identified and removed from the structure as well as any unwanted ligands such as solvents (e.g. $[NaL_3(H_2O)_3]^+$, where L is a monodentate ligand). Elements which the user wishes to exclude from the search hits can be defined in the following line. Elements must be defined using their atomic symbol.

```
    not_elements = ["Br", "Cl", "I", "Li", "Na", "K", "Ca", "Mg", "Be"]
```

It is also useful to be able to set a limit for the molecular weight of returned hit structures. For example, many entries in the CSD contain motifs such as boron cages, polymeric structures or metal-organic frameworks. Such entries are not useful as ligands in organometallic catalysis and must therefore be filtered out from the search results. The easiest way to achieve this is by setting a maximum molecular weight. Commonly used large phosphorous-based ligands such as XPhos, XantPhos and BINAP have a molecular weight of 476 Da, 578 Da and 622 Da respectively. Higher molecular weight structures are often harder to synthesise, are bulkier and may have issues with solubility. Structures exceeding approx. 700 Da are likely to be too large to be used as a ligand therefore, we can set a maximum limit to filter out these structures. The value of the molecular weight limit will be dependent on the type of ligand being searched for, e.g. smaller organic ligands for base metals or bulky phosphorous ligands for precious metals. Therefore, the molecular weight limit can be applied/adjusted using the `-w` keyword when running the search.

2.2.3.2.4 Identification of Coordinating Atoms

In order for the 3D structures from the CSD to be used for the generation of organometallic complexes, the atom indexes for the coordinating atoms of each hit structure need to be extracted and saved to a file for easy access. As the coordinating atoms in each ligand are already defined in the catalophore, the location of the features can be used to locate the coordinating atoms and extract their indexes. Each feature in **CatSD** contains the *'catsd_-'* prefix which can be used to identify the correct features with the CSD-PythonAPI and extract the XYZ coordinates of each feature point in a hit. These coordinates are unique for each hit and can then be compared with the XYZ coordinates of the hit structure to locate the base atom that the feature point matched and extract the atom index.

Due to differences in decimal place value due to differences in XYZ coordinates between the CSD entry and the CSD-CrossMiner hit and because of how computers perform and store floating point numbers, directly matching the value of the coordinates between the feature point and the base atom is not possible without treatment of the values. For example an error of 0.0001 in an x value of 1.9995 rounds up to 2.000 whereas the value it is compared to 1.9994 rounds to 1.999 and does not match. These errors can propagate forward making matching identical values difficult. Therefore, all coordinates, both from the 3D structure and the feature points, are rounded down to one decimal place and compared with a tolerance of 0.1. If all three coordinates, x, y and z are within the tolerance the atom is matched and the atom index extracted.

For *linear_nb* features the matched atom is hydrogen. In this case, as the hydrogen atom is deprotonated upon coordination to the metal centre, the atom index of the atom it is bonded to is required instead. The bonds that the hydrogen forms are retrieved from the *.mol* CSD entry and the atom index of the bonded atom extracted instead.

2.2.3.2.5 Duplicate Removal and Structure Treatment

In the CSD there can be several entries per chemical structure, including different entries and different salts. While different entries are easy to differentiate because they possess the same CSD identifier with a different number as a suffix (e.g. **ZZZLWW01**, **ZZZLWW02**, **ZZZLWW03**), different salts possess completely different CSD identifiers. Therefore, the comparison of CSD identifiers is insufficient to remove duplicates.

To remove duplicate structures SMILES matching is used to identify identical structures. As CSD-CrossMiner returns a hit object for the component of the entry matched in the search, not the full CSD entry, this means that the salts are not included in the hit SMILES string. Therefore, matching via SMILES string is a straightforward easily accessible way to locate and remove duplicate structures. If the structure of a new hit matches one that has already been processed it is discarded.

Before the *.mol* 3D structure files are saved for each hit, the structures are cleaned up to ensure they are in the correct format and all hydrogens are present for analysis or generation of complexes. All unknown bond types are assigned, all missing hydrogens are added and formal charges are set to aid the calculation of complex charges later.

2.2.3.2.6 Search Output

The results of a search are saved in the *'name.csv'* file. This file contains all of the hit structures, as their CSD Identifiers and important information in the following format.

```
CSD_Identifier, Index, Chemical Name, Structure File, Coord Atoms,
Freq, rmsd
```

`Index` is a unique suffix for each CSD Identifier which is required when duplicates are not removed, to distinguish between different crystal structures or coordination modes within the same hit. `Chemical Name` is the chemical name of the hit structure. `Structure File` is the name of the *.mol* file that is saved from the search. `Coord Atoms` are the coordinating atom indexes identified from the search and is used to generate molSimplify input files. `Freq` is the frequency for the ligand to be used in the molSimplify input files. For example, if you are searching for ligands for a trigonal planar complex with one fixed ligand, a bidentate search will have a frequency of 1 and a monodentate search will have a frequency of 2 to fill the remaining two coordination sites. `rmsd` is the root mean squared deviation between the hit structure and the catalophore.

All hit structures are saved locally in the 3D *.mol* format. The *.mol* format is used as it is a common file format for most computational chemistry programs and can be used with the molSimplify python package for building organometallic complexes for exploration and prediction.[96] A *.csv* file containing all of the SMILES strings for each hit is also saved for hit analysis or can also be used as an input for molSimplify, although this is not recommended due to frequent errors converting SMILES to 3D, often giving incorrect or unusable complexes.

The script also produces a molSimplify *.dict* file containing all of the relevant data required to use the 3D structures to generate organometallic complexes using molSimplify. The file contains the following information in the format:

```
CSD_Identifier, 3D structure file name, CSD_Identifier_1,
coordinating atom indexes, "build custom custom", "BA", formal charge
```

Values in quotes are fixed string values. `CSD_Identifier` is the CSD Identifier. `3D structure file name` is the name of the ligand *.mol* file saved from the search. `CSD_Identifier_1` is a unique name for each ligand and cannot be the same as `CSD_Identifier`. The same suffix is used as in the *.csv* `Index`. By default `_1` is used as a suffix if there are no duplicate structures. If duplicate removal is turned off the suffix `_X`, where X is an integer, is used for structures with the same CSD Identifier. `coordinating atom indexes` are the atom indexes of the coordinating atoms extracted from the CSD-CrossMiner search and are used to form the bonds between the ligand and the metal centre. `"build custom custom"` tells molSimplify that the ligand is used to build a custom complex. `"BA"` is the type of force field optimization to use by default when using the ligand to build a complex. `formal charge` is the charge of the ligand and is used to calculate the total charge of the output complex. For more information on how to use these files to generate structures using molSimplify see **Section 3.2.5.1**.

2.3 Conclusions

In conclusion, a structure-based design methodology was developed using CSD-CrossMiner and the CSD-PythonAPI to enable the searching of the Cambridge Structural Database to identify potential ligands for organometallic catalysis. A feature database, **CatSD** was developed, based on the CSD, containing structural features related to catalysis. **CatSD** enables the identification of key structural features required for catalysis, such as coordinating atom geometries, specific functional groups and other molecular motifs such as ring systems or hydrogen bonds. The ability to define the location of substrates to prevent steric clashes with the ligand(s) is also outlined.

The application of **CatSD** for searching the CSD, via the generation of a catalophore search query, using the CSD-PythonAPI was outlined using a simple example. The search script as well as its functionality was outlined to enable users to modify and use it for their own applications.

The limitations of the approach were also outlined, including being limited to organic structures and the inability to use the CSD-CrossMiner interface to conduct searches due to the requirement for additional information. The ability to generate organometallic structures from the identified 3D structures was also introduced and is explored further in **Chapter 3** and **Chapter 4**.

2.A Appendix

All python scripts used in this chapter as well as example catalophores and the CatSD feature definitions are available at `https://github.com/MarcS18/Thesis_ESI`.

Chapter 3: High Throughput Computational Workflow for Organometallic Ligand Screening

3.1 Introduction

There has long been an interest in developing low-cost computational methods to model transition-metal complexes. Computational chemistry aims to determine the electronic structure of a chemical system in order to model reactivity or predict molecular properties. The output of which can be used in cheminformatics models. The challenge of modelling transition-metal complex is their diverse range of bonding, spin states and oxidation states. This section describes both the developments and applications of a wide range of commonly available computational modelling methods, of ranging cost and accuracy, in transition-metal chemistry and their applicability in high-throughput computational screening.

3.1.1 Electronic Structure Methods

3.1.1.1 Force Fields

Affordable but reasonably accurate tools for structure generation and property prediction are useful for efficient or high-throughput computational workflows. Force fields, while popular in biochemistry and the modelling of proteins, have decreased in popularity in recent years in transition metal chemistry. The majority of force fields were developed with main group or protein chemistry in mind. However, several force fields were developed for transition metal chemistry.[97–99] The best performance is usually achieved by focusing on a subset of properties such as structure, spectra or materials.[23]

The most widely used force field for transition metals is the universal force field (UFF).[98] The universal force field uses a set of rules based on the mixing of elements and on Badger's rules for relating bond length and bond strength, enabling the generation of parameters for a large number of elements. Average errors for transition metal complexes using UFF are around 0.05 Å to 0.10 Å.[100,101] As a single oxidation state and spin state is used in parameter generation for each metal, caution should be used when applying UFF to alternative oxidation and spin states, especially in open-shell transition metal complexes.[98,101] For ligands that involve backbonding, metal-ligand bond orders must be adjusted to avoid under or overestimating bond lengths.[100]

3.1.1.2 Transition State Force Fields

Force fields are generally parameterised for metals in their equilibrium structures and therefore poor at describing transition states. Transition states are crucial for the design of transition metal catalysts and therefore to model reaction mechanisms using a force field, an alternative approach is required. The transition state force field (TSFF) approach was developed by Houk and co-workers and popularised by Norrby and co-workers as 'quantum to molecular mechanics' (Q2MM).[102,103] The TSFF approach uses a force field that has been fit on the hessian of a known transition state from quantum mechanical data (i.e. DFT). The negative eigenvalue of the hessian is inverted so that the minimisation algorithm optimises the structure to a transition state.

The TSFF approach can be applied to screen a large number of ligands. However, the main limitation is the difficulty in accurately parameterising the force field and modelling the change in steric contributions along the reaction coordinate. Q2MM has been shown to be effective in identifying small energy differences between ligands required to predict enantiomeric excess as well as identifying errors in experimental values.[104,105]

3.1.1.3 Semi-Empirical Methods

Semi-empirical quantum mechanical (SQM) methods are designed to bridge the gap between fast force field (FF) molecular mechanics methods and ab initio quantum mechanical methods (QM). SQM methods combine quantum mechanics and parameters derived from experimental data. Semi-empirical methods are faster by more than two orders of magnitude than full quantum mechanical methods and are therefore useful for initial geometry guesses and large datasets where high accuracy is not required.[106] The main disadvantage of SQM methods is that it only works for molecules within the parameter space. OM2 is used extensively for the study of excited states.[107,108] For the computation of ground state structures and energies, PM6 is favoured as it covers a large proportion of the periodic table (70 elements).[109,110] The accuracy of PM7 compared to DFT for the reproduction of conformational energies of transition metal complexes was recently assessed.[111] PM7 was shown to give large potential energy surface discrepancies from DFT, related to distortion of the coordination centre geometry and the false coordination of some atoms to the metal centre. Application of PM6/PM7 methods to transition metal complexes should be carefully examined for each system, especially where the system of interest has limited parameters in the method's training set such as reaction intermediates and transition states.

3.1.1.4 Tight Binding Methods

Density functional tight binding (DFTB) methods have emerged over the last 20 years. DFTB uses the Kohn-Sham DFT energy and expands it based on the density fluctuation, $\delta\rho$, relative to the superposition of atomic reference densities. DFTB uses element pair-specific parameterisation, which is difficult to parameterise. Currently parameterized element pairs are only applicable to simple organic molecules and materials making the unsuitable for use on transition metal complexes. To expand the applicability of tight binding methods to a wider range of molecular systems, Grimme developed an Extended Tight Binding method (xTB) for the computation of molecular geometries, vibrational frequencies and non-covalent interaction energies.[112] The methods termed GFNn-xTB (n=0,1,2) use global and element-specific parameters for elements H-Ra.

The most recent variation of the xTB method GFN2-xTB expands on the prior GFN0-xTB and GFN1-xTB methods as well as DFTB in several key areas:[113]

- GFN2-xTB uses a minimal valence basis set of atom-centred, contracted Gaussian functions as in DFT. Polarization functions are applied to main group elements to describe hypervalent structures and hydrogen is only assigned a single 1s function.
- The energy function closely resembles the DFTB3 method but expands upon it with

the inclusion of electrostatic and exchange-correlation interactions up to the second order. This allows the treatment of hydrogen and halogen bonds with electrostatics rather than force field corrections.

- GFN2-xTB includes the D4 dispersion correction method to account for dispersion interactions.
- No element-pair-specific parameters are used, so atoms are treated separately.
- GFN2-xTB focuses on geometries over bond energies when fitting, and therefore, contains a systematic error for covalent bond energies. The application of a correction factor leads to more accurate properties compared to similar SQM methods.

Grimme and co-workers recently accessed the accuracy of the GFN2-xTB method for the optimization of 145 closed shell transition metal complexes for metals up to Hg.[114] Optimised structures were compared with high-quality hybrid DFT (TPSSh-D3(BJ)-ATM/def2-TZVPP) gas phase structures or X-ray structures. GFN2-xTB was able to reproduce metal-ligand bond length and bond angles with good accuracy with a mean absolute error (MAE) of 8.3 pm and 3.9° respectively. The universal applicability, speed and relative accuracy make the GFN2-xTB method a useful tool for the study of large organometallic complexes. Recently GFN2-xTB was applied to the automatic identification of transition states in 100 organic reactions, successfully identifying 89 out of 100 transition states.[115] The mean average errors for the reaction energy barrier were 14.9-19.2 $\mathrm{kcal\,mol^{-1}}$ compared to DFT (UB3LYP/6-31G**). Errors reduced to 5.3 $\mathrm{kcal\,mol^{-1}}$ for reactions with a DFT energy barrier <30 $\mathrm{kcal\,mol^{-1}}$.[115]

3.1.1.5 Composite Methods

Composite methods have been developed by Grimme to bridge to gap between semiempirical methods and DFT for calculating properties of large molecular systems, such as supramolecular and biomolecular complexes.[116] The original HF-3c method is designed to correct for systematic deficiencies in small basis set HF calculations by correcting for basis set superposition error with the geometric counterpoise scheme (gCP), correcting for dispersion using Grimme's D3 approach and a correction for short-range basis incompleteness. As no integrals are skipped, the HF-3c method is more costly than traditional semiempirical methods, but the results are much more robust. Grimme then introduced PBEh-3c, based on the Perdew-Burke-Ernzerhof (PBE) functional with (42%) HF exchange along with a double-ζ basis set, for the optimization of geometries and for the interaction energies of non-covalent complexes.[117] Compared to HF-3c, PBEh-3c is more computationally demanding however, yields much better geometries.

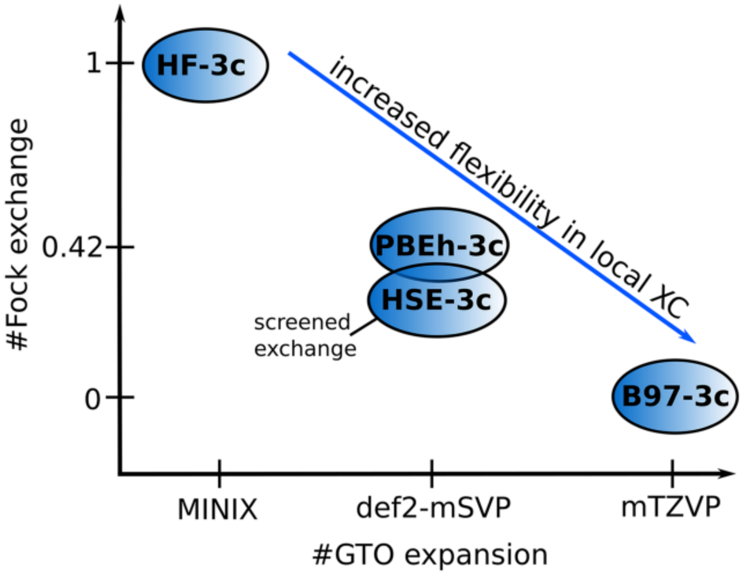


Figure 3.1: Image of the 3c composite methods according to their basis set size and amount of Fock exchange.[118]

The B97-3c method is used for the calculation of accurate thermochemistry, structures, non-covalent interactions and transition metal chemistry.[118] B97-3c is based upon the B97 GGA functional and includes the same three corrections, D3 three-body dispersion, short-range bond length correction and minor modifications to the functional along with a stripped down triple-ζ basis set, def2-mTZVP. The def2-mTZVP basis set is based on Ahlrichs def2-TZVP basis set and includes several modifications, reduced polarisation on hydrogen to decrease computational time and additional polarization functions on oxygen for a better description of strong hydrogen bonds.[118,119] The use of a GGA over hybrid DFT (which contains HF exchange) improves the treatment of electronically complicated systems such as open-shell species and transition metal complexes. The computational time for B97-3c is between HF-3c and PBEh-3c, and two to three times faster than BP86-D3/def2-TZVP (a GGA functional with the standard def2-TZVP basis set).[118] B97-3c has been shown to outperform the traditional B97-D3 GGA functional for geometries, with RMSD of 0.71 pm and 1.16 pm respectively, making it an attractive method for screening large numbers of transition metal complexes.[118]

The most recent addition to the set of composite methods is r^2SCAN-3c.[120] r^2SCAN-3c uses the r^2SCAN functional combined with a tailor-made triple-ζ basis set as well as D4 dispersion correction and geometrical counter-poise correction for London-dispersion and basis set superposition error. r^2SCAN-3c improves upon B97-3c for the prediction of main-group thermochemistry on the GMTKN55 database, at twice the cost.[120] For the reproduction of geometries however, r^2SCAN-3c performs similarly to B97-3c. For reaction and conformational energies as well as non-covalent interactions it outperforms hybrid-DFT/quadruple-ζ approaches at two to three orders of magnitude lower cost.[120]

3.1.1.6 Density Functional Theory (DFT)

Density functional theory is one of the most popular computational methods for predicting the properties of chemical systems due to its phenomenal accuracy-to-cost ratio.[121,122] Density functional theory is based on the premise that the energy of a molecule in its ground state, E_0, can be determined from the electron density, ρ. The ground state energy and molecular properties can be calculated from the ground state electron density, without using the wave function of the system (**Equation 3.1**).

$$E_0 = E_0[\rho_0] \tag{3.1}$$

DFT calculations are often very accurate and offer significantly lower costs than post-HF methods such as Moller-Plesset and coupled cluster. DFT scales to N^x $(2 < x < 3)$, where N is the number of atoms, the number of degrees of freedom in the molecule, and how close the starting structure is to the minimum.[123] Achieving a structure representing a true global minimum is often a time-consuming process as molecules can become trapped in a local minimum on the potential energy surface. Initial conformer screening commonly uses molecular mechanics due to its speed to overcome this.

The accuracy of DFT is dependent on both the functional and basis set used. Determination of a suitable functional often requires benchmarking on a relevant structure set to reproduce the property of interest. BP86 and TPSS/TPSSh functionals show excellent performance for first-row transition metal geometries while PBE0 is recommended for second and third-row transition metals.[124–126] Single-point calculations with hybrid functionals (such as TPSSh, B3LYP or PBE0) can give more reliable energies or molecular properties of transition metal compounds compared to GGA-based functionals.[127] Performance is very system-dependent and DFT functionals often fail to describe the static correlation present in open-shell transition metal compounds.[128] The covalency of metal-ligand bonds is strongly affected by the amount of Hartree-Fock exchange in the functional. GGA functionals may overestimate the covalency of bonds while hybrid functionals with high HF exchange, approximately 30% or higher, will give too ionic bonds.

The degree of complexity and therefore precision of a basis set is defined as the number of contracted gaussian-type orbitals used to represent each atomic orbital. To increase precision two or more functions can be used to describe each orbital, commonly referred to as double- or *n*-zeta basis sets.

Additional functions can be added to basis sets to describe polarisation and diffuse effects. Polarisation functions describe the asymmetric deformation of the electron cloud between atoms induced by bonding. To accommodate this, functions of higher angular momentum are included to allow for orbital combination. Diffuse functions more accurately represent the part of the atomic orbital furthest from the nuclei. These functions are required when describing anions and large, ‘soft’ molecular systems such as second and third-row transition metals. To describe transition metal systems it is recommended to use a triple-zeta basis set with polarisation functions to better describe the bonding environment, due to multiple oxidation states and complex bonding between the metal and ligands.

Transition metals possess a large core electron count and therefore, require a large number of basis functions to describe them. To reduce the number of basis functions used, those functions can be replaced by an Effective Core Potential (ECP). The ECP models the effect of the nucleus and core electrons as an average. ECPs greatly reduce the computational cost, and for atoms with Z>Kr relativistic effects can be included.

Standard DFT functionals fail to account for London dispersion. To account for this Grimme developed the D2, D3 and D4 dispersion correction methods. DFT-D3 is an atom-pairwise dispersion correction which is added to the DFT energy and gradient.

$$E_{DFT-D3} = E_{KS-DFT} + E_{disp} \tag{3.2}$$

DFT-D3 produces accurate results with DFT with little to no additional computational cost.[129,130]

3.1.1.7 Post-HF Methods

As computing power increases and the development of reduced scaling algorithms progresses, the use of correlated wave function methods will become more and more viable.[9] While single reference wave function methods such as coupled cluster display high scaling (i.e. $>O(N^5)$), they are able to achieve a chemical accuracy within 1 kcal mol^{-1} of experiment in organic chemistry.[23] While single reference wave function methods are promising they remain non-viable for high-throughput screening due to their high cost, both in computational power and time. These methods however are an effective benchmarking tool for the initial exploration of a small number of structures to access the magnitude of errors present in the lower-level methods used in high-throughput screening.

3.1.2 Computational Modelling of First Row Transition Metals

Computational investigations into reaction pathways are more complicated for paramagnetic species because the accuracy of DFT on high-multiplicity species (S>0) is dependent on the functional used.[131] Additionally, first-row transition metals often have multiple spin-states of similar energies. Since the relative energies of spin states are geometry dependant, a change in geometry during a reaction can lead to the crossing of the potential energy surfaces. Therefore, multiple energy surfaces have to be considered when studying the reaction pathways of these systems. One advantage, however, is that the complex can switch spin states to avoid high energy barriers (spin acceleration).[132,133]

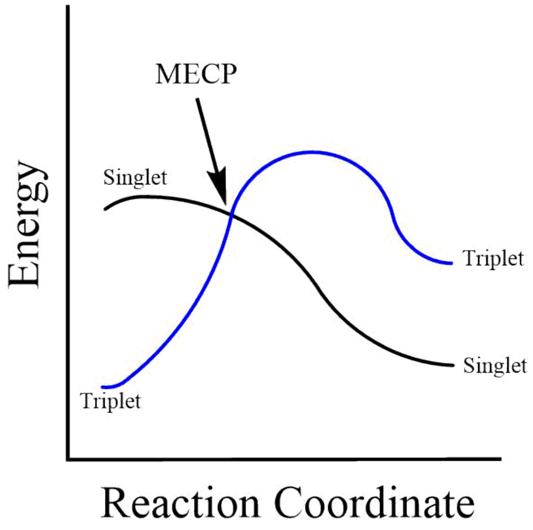


Figure 3.2: Minimum Energy Crossing Point between a singlet and triplet state.

On a three-dimensional surface such as the potential energy surface, there are multiple crossing points between spin states across the surface, the lowest of which are the minimum-energy crossing points (MECPs). If the MECP lies below both transition states there is no additional hindrance from a change in the spin state because the selection rule against spin flipping is relaxed due to strong spin-orbit coupling.[132,133] If the MECP lies above the transition state of the excited-state potential energy surface then the change in the spin state creates an additional barrier. There are currently few very examples of spin changes in transition metal catalysed reactions.[134–136]

3.2 Results & Discussion

3.2.1 Benchmarking of Computational Methods

The optimisation of transition metal complexes with DFT is often less accurate than for organic compounds due to relativistic and static correlation effects.[137,138] Most transition metal complexes used in catalysis contain large organic ligands. The relative arrangement of these ligands is determined by dispersion forces which are not accounted for in standard DFT.[139] Ensuring accurate molecular geometries is crucial for the calculation of reliable molecular properties and removing undetected errors. Benchmarking both the functional and basis set used is an important step in ensuring an accurate representation of the chemical system of interest and developing an accurate/cost-efficient method for the desired application.[140–142]

Assessing a large dataset containing both small and large ligands, where dispersion forces influence geometries, requires a method that can account for all effects. Most GGA functionals do not include dispersion correction, leading to significant errors in molecular geometries. Using functionals which are parameterized with dispersion interaction included, such as the Minnesota functionals (M06*), range-separated functionals or using Grimme's D3 dispersion correction is recommended.[129,143,144] It should be noted that the Minnesota functionals do not display the formally correct R^{-6} behaviour so the use of D3 dispersion correction is recommended.[145]

3.2.1.1 Functional Benchmark

Currently, the only method to benchmark transition metal complexes with real structures is against X-ray diffraction data. There are a few things to consider when comparing DFT-optimised structures and X-ray diffraction data. First, as atomic charge, Z, increases the x-ray scattering ability of the atom increases. It is, therefore, difficult to determine the position of light elements (small Z) in the presence of heavy elements (large Z), especially hydrogen. Secondly, geometric parameters of transition metal centres are modified by the influence of surrounding molecules compared to the gas phase and uncertainty is dependent on molecular flexibility. Typical uncertainties for transition metal structures range from 0.01-0.02 Å for metal-ligand bond lengths and 1-2° for ligand-metal-ligand bond angles, compared with <0.01 Å for organic molecules.[146] These uncertainties, however, are relatively small compared to the known overestimation of bond lengths by DFT methods.[147] Thirdly, crystal packing effects and vibrational frequencies, i.e. the difference between the DFT-minimum and the vibrationally averaged x-ray structure at the temperature of the experiment, is assumed to be small and negligible. The assumption can be justified by the fact that the basis set is fitted to match experimental data and therefore, reduces statistical error. Care should be used when describing bonds that are not well parameterised by the basis set such as metal-carbene and metal-allyl bonds. Finally, although x-ray structures are meaningful and suitable for distinguishing between suitable functionals, they are less accurate than validation against gas phase experiments such as gas electron diffraction (GED).[124]

The X-ray structures used in the benchmark were manually retrieved from the Cambridge Structural Database. Structures were chosen which met the criteria of the catalyst active state in the Ullmann-Goldberg reaction (see **Section 4.1.2.2**):

- a three-coordinate copper(I) centre
- a deprotonated N nucleophile
- two monodentate or one bidentate ligand(s) with N, O, S or P donor atom(s)

All retrieved structures contain bulky ligands and bulky electron-withdrawing nucleophiles (**Figure 3.3**), this can be assumed to be due to the low stability of the intermediate structure in the solid phase when non-bulky substituents are used. Only eleven structures meet these criteria in the CSD, one of which contains boron and therefore, is excluded due to problems with SCF convergence in DFT calculations.

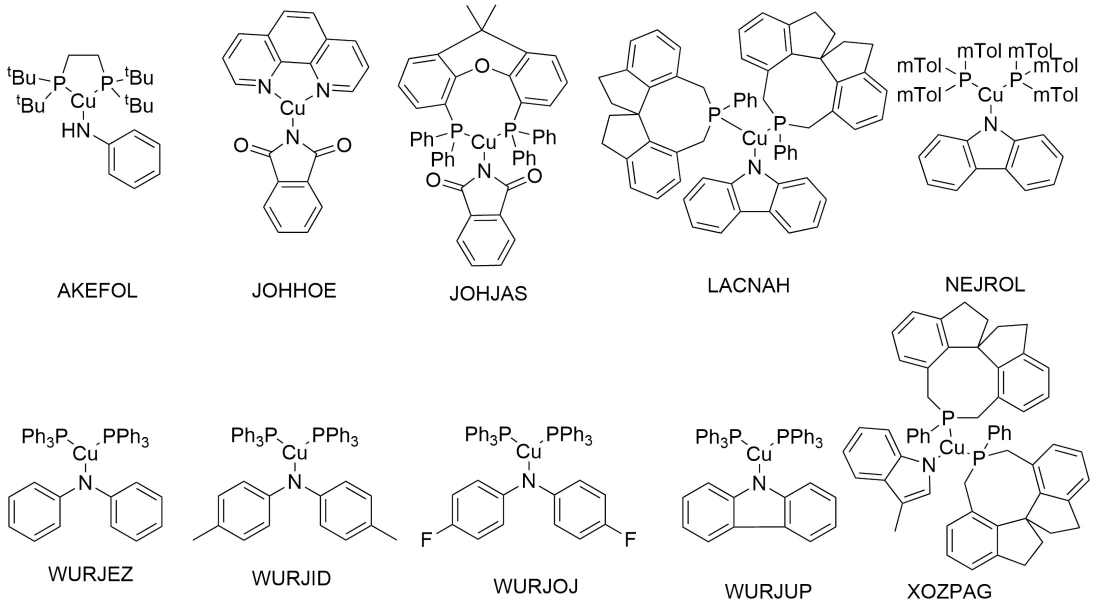


Figure 3.3: Chemical structures of the 10 available crystal structures best representing the trigonal planar active catalytic state.

Table 3.1: Copper(I) complexes used for benchmarking, their ligand types and X-ray structure accuracies.

CSD Refcode	Ligand Type	Denticity	R factor (%)	Average σ(C-C) (Å)
AKEFOL	PP	Bidentate	6.70	0.011-0.030
JOHHOE	NN	Bidentate	5.09	0.006-0.010
JOHJAS	PP	Bidentate	4.24	0.001-0.005
LACNAH	P, P	Monodentate	6.59	0.006-0.010
NEJROL	P, P	Monodentate	4.78	0.001-0.005
WURJEZ	P, P	Monodentate	4.03	0.001-0.005
WURJID	P, P	Monodentate	3.11	0.001-0.005
WURJOJ	P, P	Monodentate	4.25	0.001-0.005
WURJUP	P, P	Monodentate	3.11	0.001-0.005
XOZPAG	P, P	Monodentate	4.61	0.001-0.005

The ten structures used for benchmarking (**Table 3.1**) all contain phosphorus ligands, except one (**JOHHOE**). All structures have high C−C bond accuracy of 0.001-0.010 Å except **AKEFOL** which is higher at 0.011-0.030 Å, which is acceptable for the use of benchmarking structures as average bond length errors are expected to be in the several picometer range. Crystallographic R-factors, the discrepancy between the structure and the observed X-ray data, all lie below 7.5% so structures can be assumed to be of sufficiently high quality to represent the structure of the active catalytic state for comparison with DFT optimised geometries.

GGA, meta-GGA, hybrid and meta-hybrid functionals were chosen based on those used commonly in the literature such as B3LYP, M06, M06L and PBE0 and those that are recommended from benchmark studies, TPSS, TPSSh, BP86, wB97X-D and MPWLYP1M.[85,124,148] D3 dispersion was used for functionals for which parameters are available, using Becke-Johnson damping for all but M06, M06-L and wB97xD.[149] Calculations were also run without D3 dispersion correction to compare the effect of dispersion correction on the reproduction of complex geometries. All calculations were performed using the def2-TZVP basis set, to ensure at least triple-ζ quality on all atoms and to minimise errors due to basis set incompleteness, while still being computationally viable on large metal-containing structures. The ultrafine integration grid (99 radial shells with 590 angular points) was used to remove grid differences, as the Minnesota functionals have been shown to require the ultrafine grid for sufficient accuracy.[150]

Computational methods were compared in the reproduction of the coordination environment, using the mean absolute error (MAE) (**Equation 3.3**) in the Cu−Nu and Cu−L bond distances and L−Cu−L and L−Cu−Nu bond angles between the DFT optimised structures and X-ray structure as well as computational time for the entire benchmark data set:

$$MAE = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j>1}^{N} \left| R_{ij}(\text{DFT}) - R_{ij}(\text{X-ray}) \right| \tag{3.3}$$

where N is the number of data points and R_{ij} is the value of interest, distance or angle, between atoms i and j. All computational times are reported as total single-core CPU time.

Table 3.2: Assessment of density functionals with the def2-TZVP basis set, in terms of the mean absolute error in metal-ligand bond length (Cu−L / Å), ligand-metal-ligand bond angle (L−Cu−L, °) and single core computational time for the benchmark dataset.

	MAE		
Functional	d(Cu−L) (Å)	a(L−Cu−L) (°)	Computational Time (h)
B3LYP	0.066	4.237	4931
M06	0.034	3.413	7850
M06-L	0.020	4.127	7127
TPSSh	0.032	3.927	5435
MPWLYP1M	0.063	4.310	6508
BP86	0.030	4.157	2143
wB97xD	0.030	4.253	9220
B3LYP-D3(BJ)	0.023	3.693	6179
M06-D3(0)	0.028	3.480	11 140
M06-L-D3(0)	0.019	4.320	5271
TPSSh-D3(BJ)	0.016	3.333	6172
TPSS-D3(BJ)	0.017	3.663	3886
PBE0-D3(BJ)	0.018	3.587	5502
BP86-D3(BJ)	0.026	4.560	4036

Without D3 dispersion correction, all methods overestimate bond lengths by several picometers due to steric crowding at the metal centre caused by DFT's inability to account for

London dispersion interactions. Increased steric interactions at the metal cause elongation of the metal-ligand bond resulting in the overestimation of bond lengths. The GGA/meta-GGA functionals, BP86 and M06-L, perform better than the hybrid functionals when predicting bond lengths, and similar metal-ligand bond angles, but BP86 is much more computationally efficient taking less than half the time compared with hybrid functionals and M06-L. The most commonly used literature functionals B3LYP and M06 overestimate bond length by a large margin, 0.066 Å and 0.034 Å respectively, with M06 performing better with respect to bond angles.

Upon the inclusion of D3 dispersion correction to properly account for London dispersion forces, the meta-GGAs, TPSS-D3(BJ) and M06-L-D3(0) generally predict bond lengths better than the GGA and hybrid functionals, only beaten by TPSSh-D3(BJ). However, GGA and meta-GGAs show worse performance compared to hybrid functionals for predicting bond angles. The best-performing functional TPSShD3(BJ), a meta-GGA hybrid, shows the best performance for both bond lengths and bond angles. The inclusion of 10% HF exchange upon a meta-GGA provides a good balance between bond length and bond angle reproduction while being of similar computational cost to the other, worse-performing functionals. Functionals with lower computational costs also have lower accuracy. Overall, TPSShD3(BJ) is a good balance between computational cost and accuracy.

3.2.1.2 Basis Set Benchmark

All basis set calculations were performed with the TPSSh functional with D3 dispersion with Becke-Johnson damping (D3(BJ)) and an ultrafine integration grid. The most common literature basis sets, the Pople basis sets and LANL2DZ basis set, were used along with the def2-SVP and def2-TZVP Ahlrich basis sets. Ahlrich basis sets have defined and well-tested auxiliary basis sets available for use with the RI-J and RIJCOSX approximations in ORCA.[151] Pople and def2-SVP basis sets were applied to all atoms except Cu, which was assigned either LANL2DZ or def2-TZVP respectively. Def2-TZVP was also assigned to heteroatoms for one method denoted def2-TZVP(Het), with def2-SVP on all other atoms.

Table 3.3: Assessment of basis set combinations in terms of the mean absolute error in metal-ligand bond length (Cu−L / Å), ligand-metal-ligand bond angle (L−Cu−L / °) and single core computational time for the benchmark dataset.

Basis Set		MAE		
General	Cu	d(Cu−L) (Å)	a(L−Cu−L) (°)	Computational Time (h)
6-31G	LANL2DZ	0.064	6.123	400
6-31G(d)	LANL2DZ	0.065	6.853	670
6-31G(d,p)	LANL2DZ	0.064	6.950	692
6-31+G(d,p)	LANL2DZ	0.056	5.433	4778
def2-SVP	def2-TZVP	0.017	4.253	1414
def2-TZVP(Het)	def2-TZVP	0.017	4.183	2570
def2-TZVP	def2-TZVP	0.016	3.333	6173

The choice of basis set on the copper has a large effect on Cu-L bond distances. The use of the double-ζ LANL2DZ basis set with ECP gives MAEs of over 0.06 Å. Use of a double-ζ basis set for copper is therefore not recommended and a triple-ζ basis set (e.g. def2-TZVP) should be used. The Pople family of basis sets underperform compared to the Ahlrich family, with larger MAEs for both bond lengths and bond angles. Of the Pople basis sets the larger, 6-31+G(d,p), performs the best. All Pople basis sets show similar MAEs for bond lengths, however, the inclusion of polarisation functions on heavy atoms, gives larger bond angle errors. Further inclusion of polarisation functions on hydrogen gives slightly increased errors for bond angles and decreased errors for bond lengths. Inclusion of diffuse functions (6-31+G(d,p)) improves bond lengths and angles by up to 0.01 Å and ∼1° respectively. However, the overestimated bond lengths are still seen with the LANL2DZ basis set at 0.056 Å and a computational time of greater than six times that of the other Pople basis sets.

The Ahlrich double-ζ basis set, def2-SVP shows superior performance compared to the Pople double-ζ basis sets with a bond angle MAE over the best Pople basis set, 6-31+G(d,p) of 4.253° and 5.433° respectively, at a third of the computational cost. Bond distances appear to be primarily dictated by the basis set applied to the metal centre rather than the basis set applied to the ligands. The triple-ζ def2-TZVP provides an acceptable bond length accuracy of 0.016 Å. Changing the basis set on atoms except Cu has no effect on the Cu−L bond distance, with a difference of 0.001 Å. The inclusion of a triple-ζ basis set on the nucleophile nitrogen atoms also provides no added accuracy, while almost doubling the computational cost. The best accuracy-cost ratio basis set combination is the def2-SVP/def2-TZVP providing reasonable accuracy at just over a fifth of the computational cost of using a full triple-ζ basis set on all atoms. The final DFT method used herein is the TPSShD3(BJ) functional with the def2-SVP basis set on all atoms except Cu which use the def2-TZVP basis set.

3.2.2 Method Comparison

DFT benchmark results were compared with semi-empirical (PM6), extended tight binding and composite methods. PM6 calculations were performed in Gaussian 09, calculating the force constant on all optimisation steps. Extended tight binding calculations were performed in xtb 6.3 with the *verytight* optimization criteria. Composite methods we performed in ORCA 4.2.1 and optimised with the default convergence criteria, integration grid and slow SCF convergence (*SlowConv*). For PBEh-3c and B97-3c, the resolution of identity (RI) approximation was used for the Coulomb (J) integral (RI-J) for a significant decrease in computational time.[152] RI-J introduces a very small error which is usually smaller than basis set errors and much smaller than electronic structure method errors so can be considered negligible.[153] Zero-point energies using RI-J differ slightly from those without RI-J, but the error is systematic and cancels for relative energies.[154]

Table 3.4: Assessment of Extended Tight Binding methods in terms of the mean absolute error in metal-ligand bond length (Cu−L / Å), ligand-metal-ligand bond angle (L−Cu−L / °) and single core computational time for the benchmark dataset.

	MAE		
Method	d(Cu−L) (Å)	a(L−Cu−L) (°)	Computational Time (mins)
GFN0-xTB	0.149	13.217	53
GFN1-xTB	0.073	4.643	122
GFN2-xTB	0.029	4.657	98

GFN2-xTB is the best performer Of all three GFNn-xTB extended tight binding methods (**Table 3.4**). GFN2-xTB gives the best reproduction of bond lengths, MAE=0.029 Å, but falls behind GFN1-xTB for bond angles, albeit by a small margin of 0.014°. The large difference in bond lengths and lower computational time make it the superior extended tight-binding method. GFN0-xTB shows poor performance in all aspects with a bond length and angle error of 0.149 Å and 13.217° respectively and should not be used in these systems.

Table 3.5: Assessment of composite 3c methods in terms of the mean absolute error in metal-ligand bond length (Cu−L / Å), ligand-metal-ligand bond angle (L−Cu−L / °) and single core computational time for the benchmark dataset.

	MAE		
Method	d(Cu−L) (Å)	a(L−Cu−L) (°)	Computational Time (h)
HF-3c	0.287	29.023	145
PBEh-3c	0.035	3.247	503
B97-3c	0.013	2.307	84

Results from the composite methods are shown in **Table 3.5**. HF-3c performs very poorly with large bond length and angle errors of 0.287 Å and 29.0° respectively. A substantial improvement is made with PBEh-3c, with a bond length and angle MAE of 0.035 Å and 3.3° respectively. A surprise performer however is the B97-3c method with a reduction in MAE to 0.013 Å and 2.3° along with a greater than four times decrease in computational time compared to PBEh-3c.

The best-performing method from each class of computational method was compared to find the best cost/accuracy method for a high-throughput computational workflow.

Table 3.6: Comparison of the Mean Absolute Error of the metal-ligand bonds distance, bond angle and computational time for the PM6, GFN2-xTB, B97-3c and TPSShD3(BJ) computational methods on the benchmark dataset. Computational time is rounded to the nearest hour.

	MAE		
Method	d(Cu−L) (Å)	a(L−Cu−L) (°)	Computational Time (h)
PM6	0.066	11.267	7
GFN2-xTB	0.029	4.657	2
B97-3c	0.013	2.307	84
TPSShD3(BJ)/def2-SVP	0.017	4.253	1413
TPSShD3(BJ)/def2-TZVP	0.016	3.333	6172

Comparison of the best performers for each computational method, PM6 (semi-empirical), GFN2-xTB (extended tight binding), B97-3c (composite) and TPSShD3(BJ) (DFT), show some unexpected results. For the fast non-DFT methods, GFN2-xTB outperforms the commonly used PM6 with a greater than two times improvement in all measured criteria. GFN2-xTB however falls behind DFT in terms of accuracy, making it the superior method for pre-optimization before higher-level DFT calculations. Of the DFT-based methods, surprisingly the B97-3c composite method outperforms the best DFT functional TPSShD3(BJ), even with a full triple-ζ basis set. B97-3c is slightly better than TPSShD3(BJ) for bond lengths and shows higher accuracy in bond angles by 1° at only 1.4% of the computational cost. The use of a triple-ζ basis set along with the carefully balanced corrections included in B97-3c provides an accurate representation of the Ullmann-Goldberg reaction's active catalytic state.

The structures used for benchmarking mostly contain phosphorus-based ligands and one nitrogen ligand due to the availability of representative crystal structures. As the majority of ligands used in the Ullmann-Goldberg reaction are N, O or S donor ligands the benchmark is not fully representative of the target ligands. However, it is the best/only data available for benchmarking the active Cu(I) species. Herein, GFN2-xTB and B97-3c were taken forward as potential candidates for the electronic structure method used for high-throughput computational screening due to their impressive cost/accuracy ratios.

3.2.3 Literature Ligands: A Development Dataset

In order to develop and test a high-throughput computational workflow a suitable test dataset of ligands is required. The dataset is used to ensure that both the developed method is accurate at modelling and identifying correct organometallic intermediate and transition state complexes, and reliable, to ensure that the failure rate is kept as small as possible. A reliable process ensures that computational resources are used efficiently and the data generated is usable. For example, the calculation of activation energy requires at least two correct structures, the active intermediate and the transition state. A failure in either of the two structures makes the other unusable. Increasing the complexity of a mechanism of interest increases the number of points of failure, therefore reliability is of paramount importance.

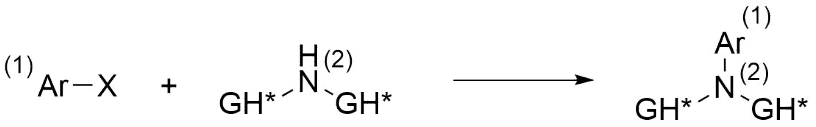
The test dataset should be made of ligand structures, ideally of those used in the literature. The use of literature ligands ensures that the workflow works for the target reaction, as intermediate structures and transition states are known to exist, and is applicable to the classes of ligands used. Extracting these ligands manually from the literature is an extremely time-consuming task, therefore a data mining approach from chemical databases was used.

Data mining is a technique for searching large-scale databases using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases, primarily used as a way to discover new meaning in data. Several chemical databases currently exist focusing on different aspects of chemical data. Chemical structure databases such as PubChem and ChemSpider focus on the curation of chemical structures and properties.[155,156] Reaxys and SciFinder focus on chemical reaction information and the Occupational Chemical Database focuses on chemical safety.[157–159]

The Reaxys database contains over 105 million organic, inorganic and organometallic compounds and reaction data from over 42 million different chemical reactions.[157] The majority of data is curated from literature and patents making it an ideal environment for data mining organometallic structures, ligands and reaction data. The raw data contained in this database, however, is not useful for comparative analysis of reactions unless extracted and curated into a new, usable format.

3.2.3.1 Dataset Curation

Chemical reaction data was retrieved for all of the available reaction conditions for C−N, C−O and C−S Ullmann-Goldberg coupling reactions. A series of searches were performed to retrieve all of the relevant data from Reaxys. Reaction type (name of a reaction) can be used to search the *Reaxys* database. The `"Ullmann condensation"` search term retrieves approximately 4000 search results, which is much lower than expected, considering it is a reasonably common reaction. Therefore, it can be assumed that the majority of the Ullmann coupling reactions present on Reaxys do not contain the `"Ullmann condensation"` reaction tag.



Scheme 8: Reaxys search term used to extract literature ligands for the Ullmann-Goldberg reaction.

Scheme 8 shows the general search term used to extract Ullmann-Goldberg reaction data from Reaxys. The requirement for an Ullmann-Goldberg coupling reaction is a halogen atom attached to an aryl/heteroaryl group which reacts with either ammonia or a primary or secondary amine/amide. Copper/Cu is defined as the catalyst, X is Cl, Br or I and Ar is any aryl or heteroaryl group. The use of the default Reaxys R group term to define R groups returns over 100,000 hit reactions. However, upon inspection, over half of these hits are reactions involving copper-palladium dual catalyst systems, Chan-Lam couplings, click chemistry and reactions where the amine is present in the aryl starting material and does not undergo a chemical transformation during the reaction. Moreover, there are no reactions containing primary or cyclic amines/amides.

The final search query uses the Reaxys GH* groups, with GH representing a general (R) group or hydrogen and * representing either cyclic or acyclic structures, ensuring all possible amine coupling partners are retrieved. To remove unwanted reaction types, several exclusion criteria were included. Palladium and Boron atoms were excluded, removing copper-palladium tandem coupling reactions and Chan-Lam couplings respectively. To exclude intramolecular reactions, unreacted aniline's and click chemistry reactions, atom mapping, denoted (n) where n is the index of atom/group, was used to map the location of the aryl group and nitrogen atom of the amine in the reactants and product. Approximately 20,000 reactions were retrieved, each of which contained multiple entries.

It should be noted that this term does not capture intramolecular Ullmann-Goldberg reactions. However, the search term covers a wide enough chemical space that any additional reactions obtained from additional searches are unlikely to contain a considerable number of unique ligands, and would require a different computational approach to model the intramolecular effects in DFT calculations. The search term was repeated for C—O and C—S coupling reaction by replacing the NH—GH* with OH and SH respectively to generate a set of raw reaction data.

3.2.3.2 Dataset Treatment and Refinement

3.2.3.2.1 Reagent Filtering

The data was exported from Reaxys in Extensible Markup Language (XML) format and the relevant data was extracted using Python. A common flaw with large chemical databases is that the data stored in the database frequently contains incorrect chemical names, incorrect spelling and is not formatted in a consistent manner. There were several issues encountered with the data structure used within the Reaxys database which had to be resolved.

```
<xf>
  <reactions>
    <reaction index="x">
      <RX> General reaction information
        <RXD.TXT> Preperation
        <RX01> Reactants
          <RX.RXRN> Reactant(s) Ref Code(s)
          <RX.RCT> Reactant Name(s)
        <RX02> Products
          <RX.PXRN> Product Ref Codes(s)
        <RX.PRO> Product Name(s)
      <RXD> Reaction conditions
        <RXD01>
          <RXD.YPRO> Product
          <RXD.YD> Yield as a percentage
          <RXD.NYD> Yield as a number
        <RXDS01>
        <RX.STG> Reaction step
          <RXD02>
            <RXD.SRCT> Reactants (not always present)
          <RXD03>
            <RXD.RGT> Reagents
          <RXD04>
            <RXD.CAT> Catalyst
          <RXD05>
            <RXD.SOL> Solvent
          <RXD.TIM> Time
          <RXD.T> Temperature
        <Citations> Citation Data
      <RY> Coordinates
        <RY.RCT> Reactant coordinates rn=reference in RX for reactant <
    RX.RXRN>
        <RY.PRO> Product coordinates rn=reference in RX for product <RX.
    PXRN>
    <\reaction>
  <\reactions>
<\xf>
```

Listing 3.1: Simplified structure of the XML file extracted from Reaxys and the data contained within each section.

Some data entries contain non-alphanumeric characters, such as subscript and italic markers, these have to be removed before data extraction otherwise data cut-off is observed.

Due to the inconsistent location of the data in the XML files, extracting the data to a useful format is non-trivial (**Listing 3.1**). As each hit (`<RX>`) has several entries (`<RXD>`), each entry has to be iterated sequentially. The Python `lxml.etree.iterparse()` parser was used to parse the XML files in an iterative manner generating a tuple of an event and its associated element. A series of if statements are used to match the tag (e.g. `RX01`) event and the associated element (`text`) is extracted and assigned to the relevant variable. For each reaction (`<RX>`) the reactants and products are assigned, and then each entry (`<RXD>`) is iterated through to retrieve all of the associated reaction data (method, yield, number of steps, reagents, catalysts, solvent, time, temperature and literature reference). After each `< RXD>` has been parsed the data is organised and manipulated to account for inconsistencies present in the Reaxys database:

1. **Catalyst:** The catalyst can be found in either the reagent (`<RXD.RGT>`), or catalyst (`<RXD.CAT>`) section of the XML file. If the catalyst is contained within `<RXD.CAT>` it can be retrieved directly. However, if the catalyst is contained within `<RXD.RGT>` a filter ('copper', 'Copper' or 'Cu') is used to find the relevant chemical name and assigned it as the catalyst.

2. **Ligand:** The Reaxys database has no specific data field for ligands, therefore, the identity of the ligand is contained within either the reagents or catalyst section. To extract the ligand a set of common reagents is used as a filter, where each reagent is checked against. If a match is found it is added to general reagents however if no match is found it is assumed to be the ligand. Common reagents include acids, bases, metal salts (excluding copper) and additives. Upon manual inspection, this method proves to be sufficient for identifying the large majority of ligands, only failing where a common reagent name is present in the ligand or when an exotic reagent is used and not caught by the filter. Ligands which are contained within the copper catalyst as a precatalyst and exotic reagents were extracted manually.

3. **Solvent:** Solvents are sometimes located within the reagents section `<RXD.RGT>`. Therefore, a solvent filter, containing common laboratory solvents, is used alongside the reagent filter to ensure the correct assignment of the solvent.

4. **Multiple catalyst entries:** In some cases entries contain multiple catalysts (`<RXD. CAT>`) data fields, one for the ligand and one for the copper species. The copper filter

is used to correctly assign the catalyst if matched, and the ligand if not matched.

5. **Reference:** All citation data is arranged into a commonly used citation format (e.g. ACS).

The treated data can then be written to the data file in the comma-separated variable (*.csv*) format and the variables cleared. After all entries (`<RXD>'s`) have been parsed the reactants and product can be cleared and the process repeated for each reaction (`<RX>`) in the XML file to give the final data file.

Before the data is ready for analysis any unwanted characters in the file are removed (e.g. '[]', ';') to make the file easier to read when analysed using commonly used spreadsheet software. Several special characters also need to be corrected to account for spelling errors and inconsistencies in the Reaxys database, for example, oxidation states ('(1)' to '(I)') and the use of grave over quotation marks (` to '). This treatment is required as these characters can cause errors when used in the command line in either Windows or Linux.

3.2.3.2.2 Entry Stripping

A key step in data analysis is the treatment of the retrieved data in order to produce a consistent dataset containing only useful data points. The parsed Reaxys data was trimmed to remove any entries that contain either no ligand or no yield, as only the ligand structures were required with their relevant yields. Manual extraction of ligands contained in precatalysts was also completed at this stage. **Table 3.7** shows the number of entries for each reaction, with a yield and an explicit ligand, extracted from Reaxys.

Table 3.7: Number of data points extracted from Reaxys for each Ullmann-Goldberg reaction class after curation.

Coupling Reaction	Curated Entries
C−N	10 728
C−O	2814
C−S	750

3.2.3.3 Retrieval of Ligand Structures

The chemical structures of the ligands were retrieved using the Chemical Identifier Resolver (CIR), an open-source cheminformatics Python module for the retrieval of chemical structures from chemical databases such as PubChem and ChemSpider.[160] Ligand structures were retrieved as SMILES strings for use with molSimplify (**Section 4.2.4.2**), as stored 3D structures in chemical databases, may not have the ligand in the correct binding conformation. The 2D nature of SMILES strings allows the generation of the 3D structure of the ligand in the correct binding conformation. Where no structure is found due to an incorrect chemical name in the Reaxys data, the ligand structure was manually retrieved. Salts and solvents were removed from the SMILES strings using the openbabel toolkit to retain only the structure with the largest molecular weight.[161] Where the ligand is the smallest component of the salt, e.g. acetate in tetrabutylphosphonium acetate, the counter-ion was removed manually. All structures were checked against the literature to ensure that the correct chemical structures were retrieved and corrected where required. SMILES strings were then stored in *.csv* format with structure name, and duplicate structures were removed to yield the final dataset (***ligands_lit_set***), containing 345 ligands (64 monodentate and 281 bidentate).

3.2.4 Reliable Optimisation of Transition States

Before a final method is chosen for high-throughput screening a reliable method for the optimisation of transition state structures must be identified. Locating and optimising transition states is a difficult and time-consuming task. Traditionally a trial and error approach is used, where an initial guess structure is generated from chemical intuition for the reaction of interest and the structure is optimised using eigenvector following algorithms. Identification of the correct transition state can take a few minutes to multiple months and a lot of computational resources depending on the method used. New methods have been developed to try and identify the transition state from the starting material and product structures. The nudged elastic band (NEB) and growing string methods take the starting materials and products as optimised structure files and identify the minimum energy path and saddle point between the two structures.[6,7,162,163] A traditional eigenvector following optimisation is then used to find the final transition state structure. These methods are popular in the solid-state and surface chemistry community but have not seen much use in molecular chemistry. This is due to the fact that in solid-state and surface chemistry one of the reactants is a static 2D object. This reduces the number of orientations possible in the starting structure if two molecules are forming a bond, making the generation of an initial guess much easier. While NEB-based methods are useful for finding a transition state they are computationally expensive due to the number of steps and gradient calculations required.

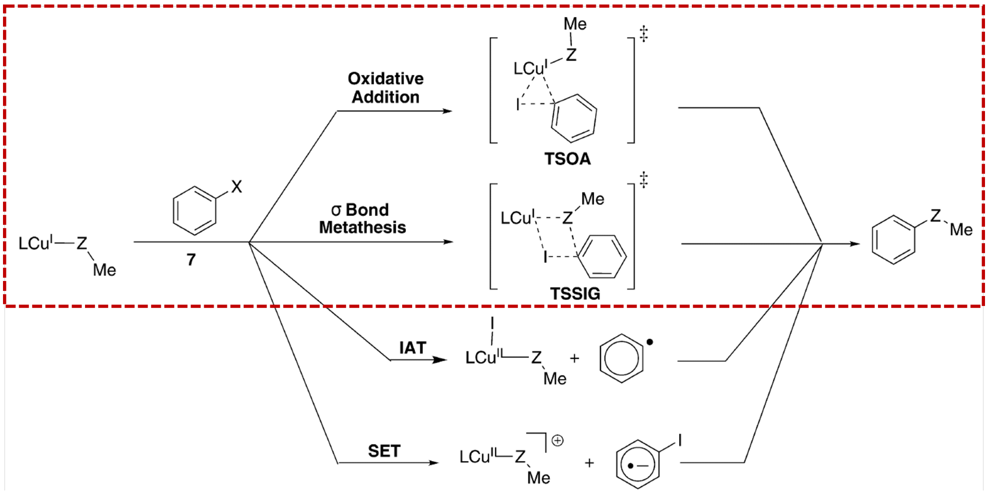


Figure 3.4: Mechanisms explored by Buchwald et al. (2010).[85]

To identify a reliable method for optimising organometallic complexes to a transition state, several transition state identification methods were tested. Nudged elastic band (NEB-TS), climbing image nudged elastic band (NEB-CI), eigenvector following from an initial guess (OptTS) and relaxed scan with subsequent eigenvector following were tested. GFN2-xTB and B97-3c were used as potential electronic methods for geometry optimisation as they were the best accuracy/cost ratio methods from the benchmarking. All methods were employed using ORCA 4.2.1 interfaced with xtb 6.3.2. Two ligands from the Buchwald mechanistic study[85] with methylamine and methanol as well as ten commonly used ligands (**Figure 3.5**) from the ***ligands_lit_set*** with piperidine were used to test each method. Iodobenzene was used as the aryl halide coupling partner in all reactions. Optimisation of two transition states, oxidative addition (**TSOA**) and sigma bond metathesis (**TSSig**) as presented by Buchwald et al. were used (**Figure 3.4**).[85]

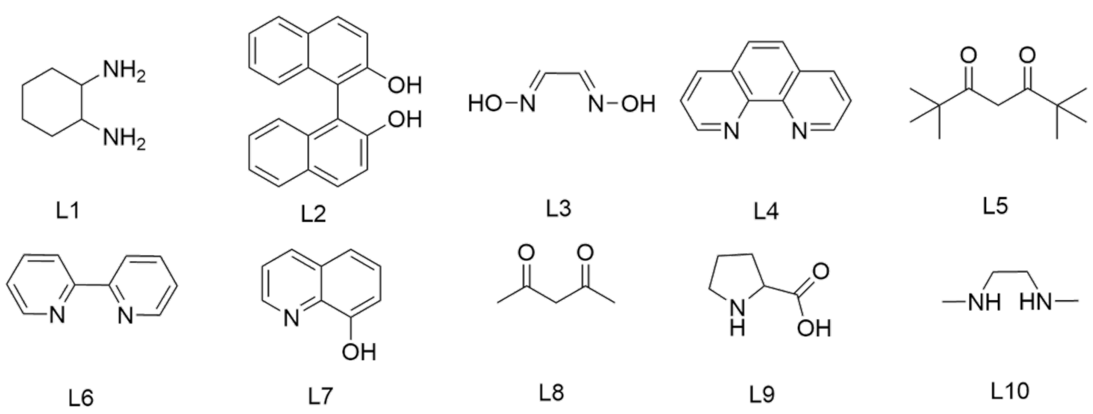


Figure 3.5: Ten ligands used for testing transition state optimisation methods.

In all cases, the NEB methods (NEB-TS and NEB-CI) were able to identify a saddle point between the starting materials and product with both GFN2-xTB and B97-3c. However, subsequent optimisation to the saddle point failed in the majority of structures often optimising to an incorrect transition state. The success of the NEB optimisations was reliant on a suitable starting orientation of the starting materials. In a high-throughput workflow, this would be difficult to determine for every single ligand. The requirement for the optimisation of both starting and final structures as well as the additional optimisation steps makes the NEB methods unsuitable for a high-throughput workflow. The same failure to optimise to a correct transition state was observed using the OptTS eigenvector following method for both GFN2-xTB and B97-3c. Including the calculation of the hessian significantly improved the success rate of correct transition state identification. Recalculation of the hessian every 20 optimisation steps was sufficient for the Buchwald ligands, successfully identifying all eight transition states. However, for the literature ligands recalculation of the hessian every 20 steps was insufficient, yielding several incorrect transition states. Reducing the number of steps until recalculation to two successfully identified all transition states across both sets of ligands. Using a hybrid hessian, including only the transition state active atoms, also proved to be ineffective at locating the correct transition state in the majority of ligands.

Recalculation of the hessian every five steps was found to be the best balance of computational time and reliability in identifying the correct transition state. The requirement for hessian calculation is likely due to the potential energy surface of the Ullmann-Goldberg reaction being very flat, especially for the oxidative addition transition state. Calculation of the hessian using B97-3c can take up to 2 days on 4 CPU cores for large ligands. This large computational requirement for these calculations makes B97-3c not viable as an optimisation method for a high-throughput workflow. In comparison, the hessian calculation at the GFN2-xTB level of theory takes approximately 20 minutes on 4 CPU cores. This reduced hessian calculation time along with good accuracy in benchmarking makes GFN2-xTB an attractive choice for high-throughput computation. Therefore, GFN2-xTB was chosen as the method for all optimisation steps for both intermediate and transition state structures as well as frequency calculations.

Another important factor in the reliable optimisation of transition states is generating a good starting structure. The starting structure must be both a good initial guess and easy to generate automatically. Transition state templating satisfies both criteria. A template of the transition state of interest is generated at the same level of theory as the optimisation method using a simple ligand. The simple ligand is then replaced with the ligand of interest, ensuring the initial guess is as close to a correct transition state as possible. This process can be easily automated with Python modules such as molSimplify.[96] To further improve the reliability of the transition state optimisation the ligand can be pre-optimised by constraining the substrates and optimising the ligand to a minimum. Pre-optimisation removes all of the imaginary frequency originating in the ligand increasing the chance that the eigenvector followed by the transition state optimisation is the one corresponding to the correct transition state.

A summary of the transition state optimisation process is described below:

1. Generate a transition state at the GFN2-xTB level of theory using a simple ligand.
2. Automatically replace the ligand using Python.
3. Constrain the substrates and pre-optimise the ligand to a minimum at the GFN2-xTB level of theory.
4. Optimise the entire structure to a transition state at the GFN2-xTB level of theory using eigenvector following.
5. Perform frequency calculation to confirm that the structure is at a minimum or has one imaginary frequency, as well as calculate thermochemistry.

3.2.4.1 Automatic Identification of Correct Transition States

Manual validation of transition state structures is not a viable approach for a high-throughput screening workflow. The vast number of ligands used (>100) and the requirement to generate a visualisation of the transition state is an extremely time-consuming process subject to human error. Automated validation of both stable and transition state structures is a requirement to make the process viable on a large scale. Automation of the process reduces both the total time and error in the analysis stage as well as frees up the chemist to interpret the data. While the validation of intermediate structures is relatively simple, requiring only the presence of zero imaginary frequencies, validation of transition states is more difficult. Transition states require only one imaginary frequency and that the imaginary frequency corresponds to the correct bond-forming/breaking process.

In order to validate the structure of the transition state a modified version of the TS vetting requirements presented by Jacobsen et al. is used.[164] This validation procedure is not based on an intrinsic reaction coordinate (IRC) calculation and therefore, reduces the total computational time and resources required. The transition state structure must meet all of the following three criteria: i) exactly one imaginary frequency of the hessian. ii) the TS active bond (bond being broken or formed) must be of an intermediate length:

$$1.7 \geq \frac{r_{ij}}{(r_i^{cov} + r_j^{cov})} > 1.0 \tag{3.4}$$

where r_{ij} is the bond length between atoms i and j and r_i^{cov} and r_j^{cov} are the covalent radii of atoms i and j. iii) the eigenvector corresponding to the imaginary frequency should have motion along one of the TS active bond stretching modes:

$$\left| v_i^{stretch} \cdot v^{ts} \right| \geq S_0 \tag{3.5}$$

where $v_i^{stretch}$ is the eigenvector of the imaginary frequency, $v_i^{stretch}$ is the unit vector of the stretching mode of bond i and S_0 is the amount of overlap between the two vectors. S_0 is a constant of default value 0.33. The value of S_0 needs to be tuned depending on the type of transition state. Transition states which are not a simple bond stretch along the TS active bonds are not well described with an S_0 value of 0.33. For example, in transition states possessing a bend-like character, commonly observed in some oxidative additions, tuning the value of S_0 is required.

3.2.5 Automated Structure Generation

In order to explore organometallic mechanistic pathways with a large number of ligands (>1000) in an efficient manner, structures must be generated automatically. Individual generation of 10,000's of structures, depending on mechanistic complexity, is not a viable method, both in time and resources. To explore a mechanistic pathway all-important intermediate structures, as well as transition states must be generated with the ligand of interest for each step. For each step, an organometallic complex must be generated with specific ligands and properties such as coordination number, charge and spin. Each organometallic complex in a step should possess very similar properties such as the oxidation state of the metal centre, spin (depending on the metal), coordination number and geometry, with the only difference being the ligands present. The type of ligand(s) present will determine the charge of the complex, which must therefore be calculated automatically. An effective approach is therefore to define a base structure for each complex of interest and then replace/add the ligand(s) automatically and calculate the charge and spin of the complex for use in computational calculations.

3.2.5.1 Generation of stable/intermediate structures

Organometallic complexes are generated in an automated manner using the molSimplify Python toolkit.[96] For each organometallic complex in a mechanistic pathway the metal centre including oxidation state and spin (in most cases), coordination number, and geometry is constant between complexes in the same step. This creates a template structure where only the ligand(s) needs to be adjusted for each structure. The ligand(s) and their frequency in the complex can then be defined for each unique ligand, the values of which depend on the ligand denticity or requirements for the final complex (additional fixed ligands). Any ligand that is required in every complex (fixed) can be included within the template. This process can be automated using Python with a suitable data source containing all of the required ligand data (3D structure/SMILES, frequency and indexes of the coordinating atoms). This allows the automated generation of molSimplify input files and subsequent automatic generation of the organometallic complexes and their respective charges and spins. MolSimplify generates the charge of the complex automatically using openbabel.[161] In cases where a ligand must be deprotonated upon coordination to the metal centre, deprotonation is done automatically using a set of deprotonation rules which use SMILES matching to match functional groups. The deprotonation rules can be customised based on the user's needs. Structures can also be optimised with a force field to clean up the structures both before and after ligand addition. An example molSimplify input file for the generation of the active catalytic species in the Ullmann-Goldberg reaction, which is a stable three-coordinate Cu(I) complex with one bidentate ligand and one deprotonated nucleophile is shown below.

```
-name AADMPY10_CuLpyr_1
-core copper
-oxstate I
-coord 3
-geometry tpl
-lig AADMPY10, pyrrolidinone
-ligocc 1, 1
-spin 1
-ff uff
-ffoption ba
-keepHs auto, False
-ligalign true
-skipANN true
```

The file contains all of the relevant information regarding the complex to be generated, including information about the metal centre, the structure of the ligands, how to deprotonate the ligands and whether to clean up the structure using a force field. Let's break down each line.

```
-name AADMPY10_CuLpyr_1
```

The `-name` line defines the name of the structure to be used in the output files. For example, the above example will output the structure as a *.xyz* file to the location `./AADMPY10_CuLpyr_1/AADMPY10_CuLpyr_1/AADMPY10_CuLpyr_1.xyz`. This value should contain all of the relevant information required to identify the complex. This provides a file structure where all of the complexes are separated allowing for easier handling of computations and analysis.

```
    -core copper
    -oxstate I
    -spin 1
```

The metal centre is defined using the above terms. `-core` states the element to be used as the metal centre. The value can either be the name of the element or the atomic symbol (e.g. copper, cu, iron, fe). The oxidation state of the metal is defined with the `-oxstate` keyword. Values use roman numerals (e.g. I, IV, V) for positive oxidation states or negative numbers (-1, -2, ...) for negative oxidation states. Finally, the spin of the metal centre is defined with the `-spin` keyword and its value is the spin multiplicity of the metal centre (e.g. 1 - singlet, 2 - doublet, 3 - triplet).

```
    -coord 3
    -geometry tpl
```

To define the geometry of the complex the following keywords are used. `-coord` defines the coordination number of the complex (number of bonds between the metal centre and ligands). `-geometry` defines the geometry of the complex (e.g. tpl - trigonal planar). For a full list of values see **Appendix 3.A.1**. Custom geometries can be used if required, please see the official molSimplify documentation for instructions.[96]

```
    -lig AADMPY10, pyrrolidinone
    -ligocc 1, 1
```

Next, the ligand(s) to be added to the metal are defined. `-lig` contains the names of the ligands to be added. The names defined here are looked up from the `ligands.dict` file located in the default molSimplify folder which is by default located at `/home/username /molSimplify/Ligands/`. This is also the same folder where the 3D structure files must be present in order for the script to work correctly. The CSD-CrossMiner search script generates a *.dict (where * is the name of the search) file containing all of the relevant values for each ligand from the CSD. The contents of which must be copied directly into the default `ligands.dict` file. 3D structure files exported from the CSD-CrossMiner search must also be copied to the `/home/username/molSimplify/Ligands/` folder.

The `-ligocc` keyword defines the frequency of each ligand in the same order they are stated in `-lig`. In this example, **AADMPY10** is a bidentate ligand and therefore only one is required for the trigonal planar complex along with one nucleophile. If **AADMPY10** was a monodentate ligand `-ligocc 2, 1` would be the correct values to include two monodentate ligands and one nucleophile to fill all coordination sites.

```
-lig AADMPY10.mol, C1CC(=O)NC1
-ligocc 1, 1
-smicat [2, 19], [4]
```

Structures can also be generated from the 3D structure files located in the same folder as the input files by defining the file name of the ligand 3D structure (e.g. AADMPY10.mol) or by SMILES string and the indexes of the coordinating atoms (e.g. [2, 19]) in the SMILES string. Atom indexes must be enclosed in square brackets for each ligand and separated by a comma using the `-smicat` keyword. The indexes of the coordinating atoms must be defined for every explicitly defined file/SMILES string. **Note:** Indexes for atoms start from 0.

```
-ff uff
-ffoption ba
```

The complex can be cleaned up using a force field at several stages during structure generation. `-ff` defines the force field to be used. For organometallic structures, only the UFF (Univeral Force Field) force field is recommended. `-ffoption` defines when the structure is optimised. **b** optimises the ligands before they are added to the metal centre. This is only recommended when using SMILES strings. **a** optimises the structure after the ligands have been added. Both options can be used together (**ba**), which will optimise the ligands before addition and the complex after all ligands are added. It is recommended to use at least the **a** option to clean up the structure, especially for bulky ligands where there is a lot of steric congestion. This ensures that any unusual structures generated during ligand addition are cleaned up before being used as a starting point for computational calculations. For example, two ligands being very close together can cause errors in ligand structures in subsequent calculations.

```
-keepHs auto, False
```

It is common that upon coordination to a metal the ligand is deprotonated. In these cases, hydrogen atoms need to be removed from the starting 3D structure of the ligand. This can be done using the `-keepHs` keyword. In order to deprotonate the ligand, for example, the active catalytic state in the Ullmann-Goldberg reaction has a deprotonated nucleophile, the `False` value can be used. To keep all hydrogen atoms use the `True` value. Custom deprotonation rules can be used for deprotonation for a large range of ligands containing a variety of functional groups. This can be achieved using the `auto` value which uses SMARTS matching to deprotonate matching functional groups. Custom deprotonation rules can be defined by altering the `self.remHsmarts` line in the `molSimplify/Classes/globalvars .py` file by replacing the list of SMARTS strings with a user-defined list. Ensure that the order of `-keepHs` values matches the ligand list in `-lig`.

```
-ligalign true
```

The `-ligalign` keyword is used to call the ligand alignment tool. This ensures that the ligands are added to the metal in order of steric bulk. This improves the structures generated as adding bulky ligands last can often lead to them not having enough space causing incorrect structures or failure of the program. Possible values are `true` and `false`.

```
-skipANN true
```

The `-skipANN` keyword is used to call the use of ANN-calculated bond lengths. This is only supported by molSimplify for specific elements and geometries. Ensure that the metal and geometry you are using are supported otherwise it can be turned off using the `true` value to save computational time.

Structures are generated automatically with the command:

```
molsimplify -i {input_file}
```

To generate all structures instead of just one, bash can be used to loop over all molSimplify input files:

```
for f in *.inp; do
    molsimplify -i {f};
done
```

The calculated charge and spin values for the complex are output to the `terachem_input` file and are important for the generation of computational input files.

3.2.5.2 Generation of transition state structures

Transition state structures are much harder to generate from scratch due to the non-standard bond lengths and bond angles present in the transition state. To generate a good starting structure for a transition state calculation the ligand replacement tool in molSimplify is used to replace the ligand in a transition state template with the ligands retrieved from CSD-CrossMiner, or from a SMILES string or other 3D structure file (e.g. *.xyz* or *.mol*).[96]

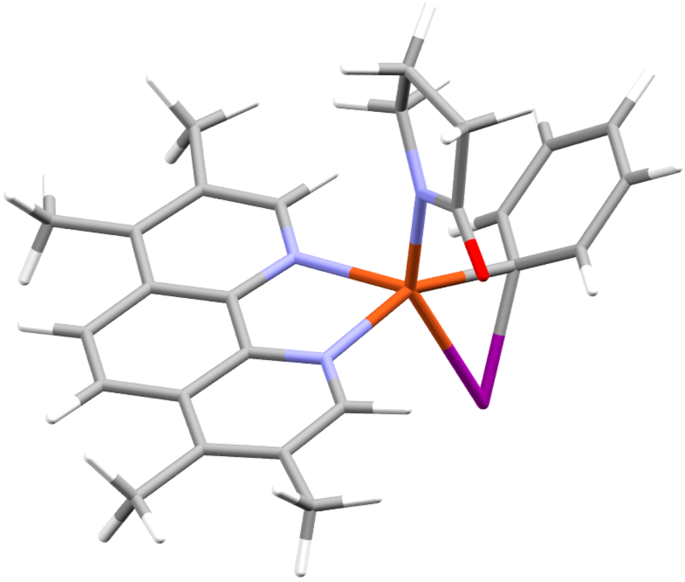


Figure 3.6: Transition state core for the oxidative-addition transition state of the Ullmann-Goldberg reaction. 2-pyrrolidinone and iodobenzene are used as substrates.

Structures are built from a template structure that uses a 'simple' and 'common' ligand as a base. The template should be an optimised transition state of the transition state of interest. This structure should be optimised at the same level of theory as the computational calculations to be employed for the entire ligand dataset. An example 'core' is shown in **Figure 3.6** for the oxidative-addition transition state for 2-pyrrolidinone and iodobenzene using 3,4,7,8-tetramethyl-1,10-phenanthroline (**TMPHEN**) as the base ligand. An example input file for automated ligand replacement with molSimplify is shown below.

```
-name AADMPY10_TSOA_pyr_1

-core tsoa_pyr

-oxstate 0

-spin 1

-replig true

-lig AADMPY10

-ligocc 1

-ccatoms 6,15

-ligloc true

-ligalign true

-keepHs auto

-ffoption c

-skipANN true
```

The input file has several differences compared to the stable intermediates. Let's break down each line.

```
-name AADMPY10_TSOA_pyr_1

-ligalign true

-skipANN true
```

The above lines are the same as the stable/intermediate structure generation, see **Section 3.2.5.1** for an explanation of these keywords.

```
-core tsoa_pyr

-oxstate 0

-spin 1
```

Instead of a metal centre `-core` defines the name of the transition state template. This is the structure of the optimised transition state to be used as a template for ligand replacement. The structure file should be placed in the `/home/username/molSimplify/Cores/` folder and an entry added to the `cores.dict` file containing the following information in the following format: `"alias":"name of the XYZ file","indexes of the coordinating atoms", "maximum denticity"`. For the oxidative-addition transition state for 2-pyrrolidinone using 3,4,7,8-tetramethyl-1,10-phenanthroline as the templating ligand, the entry is "tsoa_pyr:TSOA_PYR.xyz,6 15,6". Where atom indexes 6 and 15 are the indexes of the two **TMPHEN** nitrogen atoms and the core has a maximum denticity of 6. The -oxstate keyword is the oxidation state of the entire core. In this example the metal centre is copper(I) and is bound to a deprotonated nucleophile of charge, -1, and a neutral ligand so the core has an oxidation state of 0. -spin is the spin of the core. In most cases, this will be the same as the stable/intermediate complexes.

```
    -replig true
    -lig AADMPY10
    -ligocc 1
```

In order to enable ligand replacement the -replig true command has to be added. The ligand to be added is defined as -lig "ligand". As with the stable complexes the ligand can be defined as a 3D structure from the ligands.dict file or be defined as a 3D structure file in the current folder or a SMILES string. The ligand coordinating atoms are taken directly from the ligands.dict file. When using a structure not defined in ligands.dict, e.g. a SMILES string, the -smicat keyword should be used to define the indexes of the coordinating atoms as described for intermediate structures (see **Section 3.2.5.1**). The number of each ligand is defined using -ligocc "frequency".

```
    -ccatoms 6,15
```

-ccatoms defines the atom indexes of the atoms in the ligand to be replaced in the core which coordinates to the metal centre. -ccatoms 6,15 refers to the two nitrogen atoms in the **TMPHEN** ligand in the core.

```
    -ligloc true
```

The -ligloc keyword enforces ligand location. This ensures that the ligand is placed in the correct position around the metal centre.

```
    -keepHs auto
```

As in the stable complexes -keepHs is used to deprotonate the ligand structures. As only the ligand is added to the structure only auto need to be used if using custom deprotonation rules. true and false can be used if no deprotonation or forced deprotonation is required respectively.

```
    -ffoption c
```

In order to maintain the transition state mode in structure generation both force field options **b** and **a** cannot be used as they will optimise the structure to a minimum and remove the transition state mode. A new `-ffoption` **c** has been implemented to fix this. **c** stands for core-constrained and freezes all the atoms in the core, resulting in only the ligand being optimised. As the core is already optimised at the desired level of theory using a force field on this part of the structure will move the structure away from the minimum. Therefore the core-constrained optimization both maintains the transition state mode and structure of the substrates as well as minimises the ligand structure. **c** can be used alongside **b** as **b** does not optimise the core, only the ligand before addition.

Table 3.8: Success rate for different force field optimization methods for transition state generation for the Ullmann-Goldberg reaction, (TSOA: oxidative-addition, TSSig: sigma-metathesis) for a set of 345 literature ligands using SMILES strings for the ligands.

Success Rates	Before (**b**)	Core (**c**)	Before + Core (**bc**)	No Force Field
TSOA	71%	87%	86%	68%
TSSig	82%	93%	93%	79%
Total	76%	90%	90%	74%

Table 3.8 shows the success rate for each force field option for the generation of two transition states (TSOA and TSSig) for the Ullmann-Goldberg reaction using 345 literature ligands. Ligands were added to the structure via SMILES string. Using core-constrained optimization improves the success rate of structure generation by ∼15% compared to using no force field and before optimization. Using both before and core-constrained optimization offers no advantage compared to just core-constrained force field optimization but comes at an additional computational cost from the additional force field optimisation step.

3.2.5.3 Correction of Coordinating Atom Indexes

Due to the deprotonation of some ligands during complex generation, the atom indexes of the coordinating atoms in the ligand may be different from those extracted from CSD-CrossMiner. This is due to the removal of the hydrogens in the atom lists. The location at which the hydrogens were present in the atom list determines whether correction of these indexes is required (**Figure 3.7**). Correction of the atom indexes is required for the correct analysis of these atoms when extracting specific properties from computational output files.

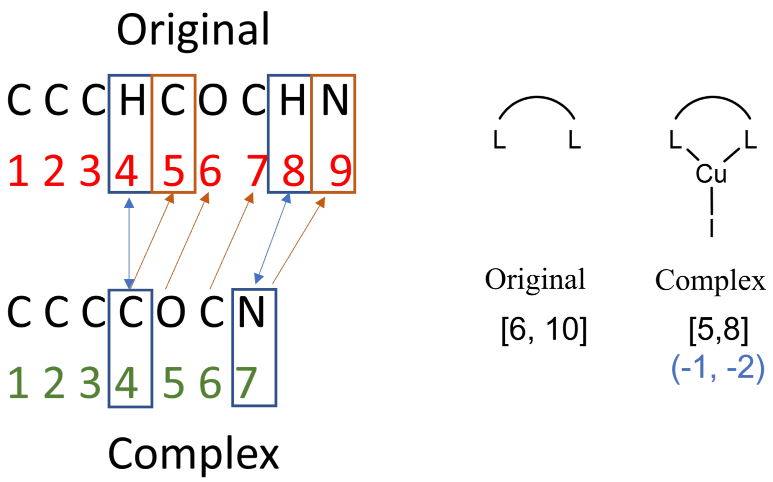


Figure 3.7: Example of the shift in coordinating atom index upon complex generation due to deprotonation.

To correct the indexes of these atoms the following method is employed:

1. The atom lists for the **CuLI** and ligand is read from their respective *.xyz*/*.mol* files.

2. The Cu and I atoms are removed from the **CuLI** complex atom list.

3. The two lists of atoms are compared.

The complex atom list is iterated through and the atomic symbol is compared to the free ligand atom list. When a miss-match is found the index (n) is recorded and then compared against the next item (n + the number of miss-matched items) in the list. If the index of the miss-matched item is greater than the index of the coordinating atoms no adjustments are needed. If the index of the miss-matched item is before the index of the coordinating atom a hydrogen has been removed. Therefore, the coordinating atom index is adjusted based on the number of miss-matched indexes lower than the coordinating atom index. This process fails if a hydrogen has been removed from a section of an atom list containing multiple hydrogen atoms (**Figure 3.8**). In this case, the following process is employed:

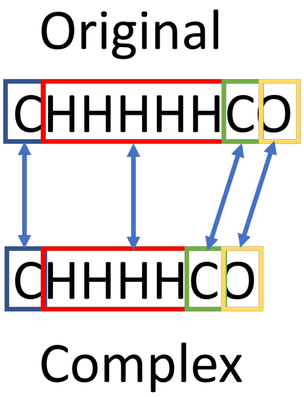


Figure 3.8: Example of the shift in coordinating atom index upon complex generation due to deprotonation where both hydrogens are in the same block.

For each atom at location (n) in the list, the list is propagated forward to find the length of the section in the list with the same repeating atomic symbol. This process is done for both structures. The lengths of the repeating sections are then compared, if they are not equal then the deprotonated hydrogen was present in that section of the atom list. The starting index of the section containing the removed hydrogen is then compared to the index of the coordinating atoms. The index is then adjusted using the same comparison method.

3.2.6 Calculation of Activation Energies

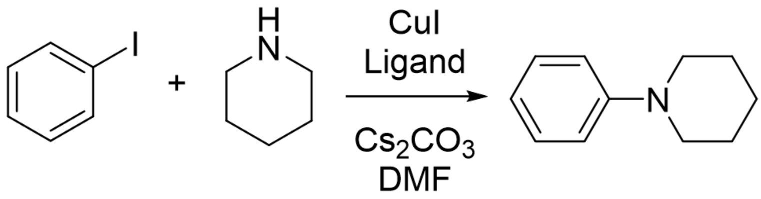


Figure 3.9: Reaction studied to analyse the computational workflow.

Activation energies for all ligands in the ***ligands_lit_set*** were calculated using the developed workflow. Activation energies were calculated for the oxidative addition (**TSOA**) and sigma metathesis (**TSSig**) pathways for the Ullmann-Goldberg reaction between piperidine and iodobenzene. Ligand structures were supplied via SMILES string. DMF was used as the solvent and caesium carbonate as the base.

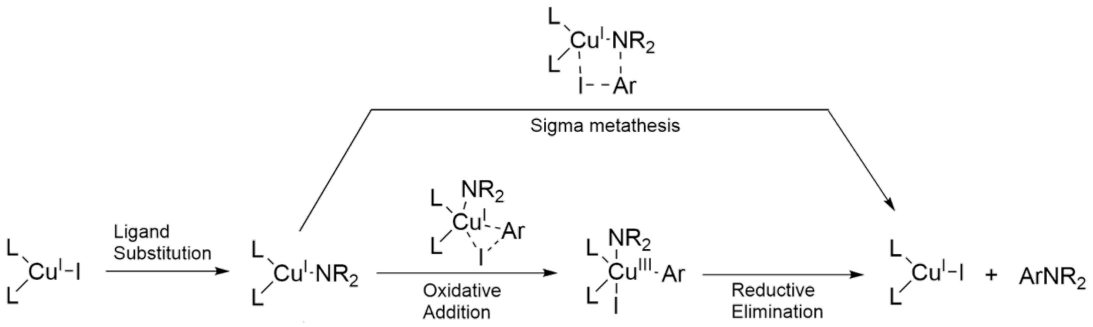


Figure 3.10: TSOA and **TSSig** pathways for the Ullmann-Goldberg reaction.

All structures in the mechanisms were generated automatically as described in **Section 3.2.5**. For the intermediate complexes, **CuLI** and **CuLpip** a trigonal planar geometry was used with a Cu centre with an oxidation state of I and spin 0. Iodide and piperidine ligands were supplied as SMILES strings. Forced deprotonation was used for piperidine. Complexes were optimised before and after generation using the UFF force field. Structures were optimised to a minimum and frequencies were calculated at the GFN2-xTB level of theory. Single point energies were calculated at the B97-3c level of theory.

For transition states, a transition state core for both the **TSOA** and **TSSig** transitions state was generated at the GFN2-xTB level of theory using **TMPHEN** as a base ligand. The optimised transition states were used as cores for structure generation. Geometry was set to trigonal bipyramidal for **TSOA** and trigonal planar for **TSSig**. Spin and charge were set to 0. Ligands were replaced automatically by defining the atom indexes of the coordinating nitrogen atoms in the **TMPHEN** ligand. The initial ligand structure was optimised with the Universal Force Field using the core-constrained optimisation method. Structures were pre-optimised using GFN2-xTB by constraining the Cu, C and I atoms involved in the bond breaking and forming step for **TSOA** and the Cu, N, C and I atoms involved in the bond breaking and forming step in **TSSig**. The resulting structures were optimised to a transition state in ORCA 4.2.1 using GFN2-xTB. Frequencies were calculated at the GFN2-xTB level of theory and single point energies at the B97-3c level of theory.

Additives (e.g. caesium carbonate and ions) were optimised to a minimum and frequencies were calculated at the GFN2-xTB level of theory. Single point energies were calculated at the B97-3c level of theory.

All single point energies are converted to the Gibbs free energy by the addition of the correction term from the frequency calculation:[165]

$$G^0 = E_{el}(\text{B97-3c}) + G_{correction}(\text{GFN2-xTB}) \tag{3.6}$$

Gibbs's free energies for additives were stored in a database. Activation energies are calculated as the difference in the sum of the Gibbs free energies of the components in each step of the reaction:

$$E_A = \sum_{1}^{n} G_{products} - \sum_{1}^{n} G_{reactants} \tag{3.7}$$

The activation energy of each pathway was calculated as the difference between the Gibbs free energy of the transition state and the lowest energy intermediate structure. Relative energies were calculated relative to the **CuLI** complex. All calculations were run on the ARC3 supercomputer at the University of Leeds. All calculations used 4 CPU cores and 4GB RAM. Resulting relative and activation energies were stored as the ***data_lit_set*** dataset.

3.2.7 Evaluation of Accuracy

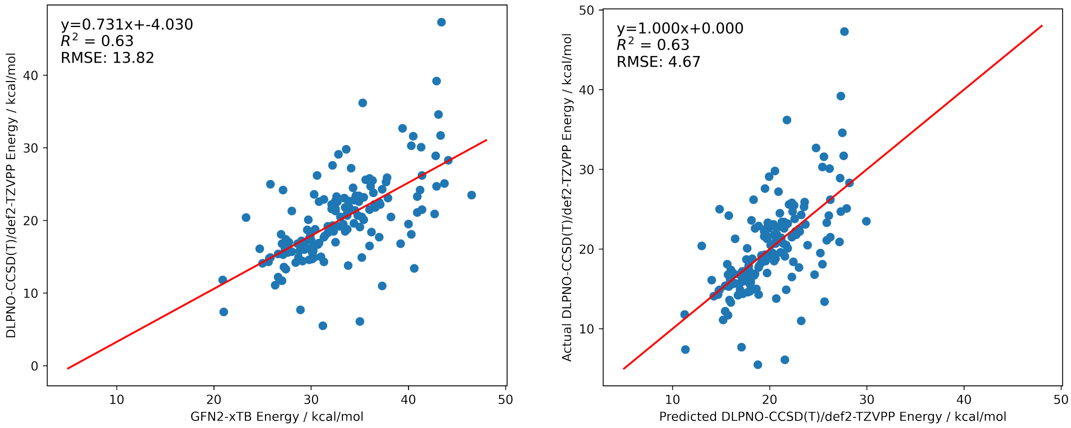
3.2.7.1 Activation Energy Benchmark

In general, DFT is known to reproduce geometries and frequencies with reasonable quality for its low cost, but energies are a weak point. Calculated energies are generally poor if the system studied is outside the training set of the functional used, which normally does not include transition states or atypical bonding situations. To assess the accuracy of the calculated activation energies, the GFN2-xTB and B97-3c relative energies for 100 random ligands from the ***ligands_lit_set*** were compared against the 'gold standard' domain-based local pair natural orbitals - coupled cluster singles doubles and perturbative triples (DLPNO-CCSD(T)) wavefunction method for the ligand exchange step, oxidative addition transition state and sigma metathesis transition state.[9] Correlated methods such as CCSD(T) are unparametrised, based on real physical principles and should truly reproduce experimental results. Both energy calculations use the same GFN2-xTB optimised structures. Coupled cluster energies were calculated at the DLPNO-CCSD(T)/def2-TZVPP level of theory and compared to the GFN2-xTB and B97-3c energies. The def2-TZVPP basis set was chosen to ensure sufficient basis set completeness as coupled cluster methods are much more sensitive to basis set completeness compared to DFT, while not requiring excessive computational resources.[166]

In order to transform the DLPNO-CCSD(T) electronic energy (E_{el}) into a true G^0, vibrational corrections must be included, and calculated using either DFT or xTB. To account for solvation, G_{CDS} and G_{ENP}, from an energy or frequency calculation computed with DFT or xTB with the desired solvent is added to the electronic energy to obtain G_{solv}:

$$G_{solv} = E_{el}(\text{DLPNO-CCSD(T)}) + G_{correction}(\text{DFT}) + G_{CDS}(\text{DFT}) + G_{ENP}(\text{DFT}) \tag{3.8}$$

where G_{solv} is the Gibbs free energy in the solvent, E_{el}(DLPNO-CCSD(T)) is the zero-point energy calculated at the DLPNO-CCSD(T) level, $G_{correction}$(DFT) is the energy required to transform E_{el} into G^0 at a given temperature and pressure and is only dependent on geometry and frequencies so can be calculated at a lower level of theory such as DFT or xTB, G_{CDS} is the cavity term and G_{ENP} is the entropy term from the interaction of the medium and the molecular surface charges. Confirmation of the reliability of the DLPNO-CCSD(T) zero-point electronic energy is verified by ensuring the T1 diagnostic is less than 0.02 to confirm the reliability of the orbitals and T2 amplitudes of less than 1 to check for multi-reference character. The solvent model based on density (SMD) implicit solvent model was used for B97-3c and DLPNO-CCSD(T), using DMF as the solvent. The generalised Born model with surface area contributions (GBSA) implicit solvation model was used for GFN2-xTB, using DMF as the solvent. All relative energies were calculated relative to the **CuLI** starting complex. The activation energy was calculated as the difference between the Gibbs free energy of the transition state and the lowest energy intermediate structure.

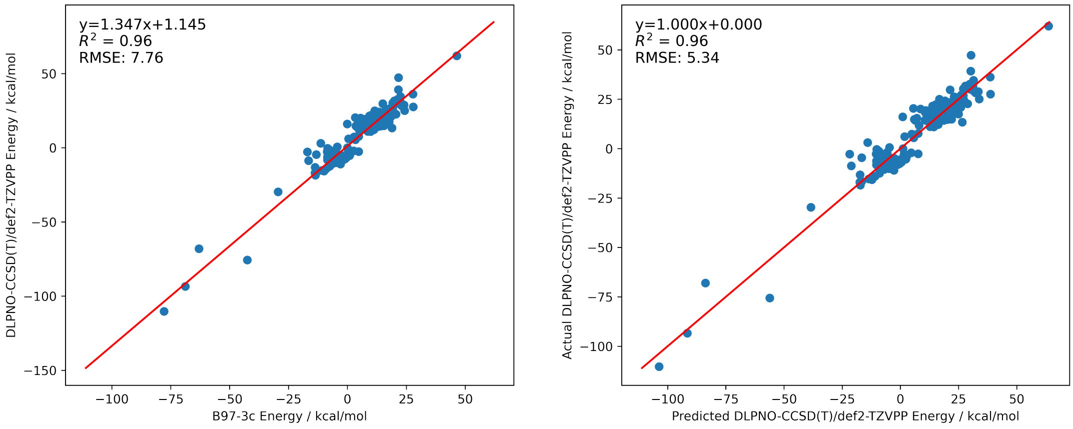


(a) Raw Values

(b) Scaled

Figure 3.11: Comparison of activation energies for GFN2-xTB vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, including ligand exchange, TSOA and TSSig.

Figure 3.11 shows the the correlation between GFN2-xTB and DLPNO-CCSD(T)/def2-TZVPP relative energies. GFN2-xTB energies correlate poorly to coupled cluster with an R^2 value of 0.63 and RMSE of 13.82 kcal mol^{-1}. When the relative GFN2-xTB energy is scaled to the DLPNO-CCSD(T)/def2-TZVPP energy using the equation of the line the RMSE decreases to 4.67 kcal mol^{-1}, however the R^2 value remains poor. Therefore, GFN2-xTB should not to used to calculate the Gibbs free energy of the structures in the Ullmann-Goldberg reaction.

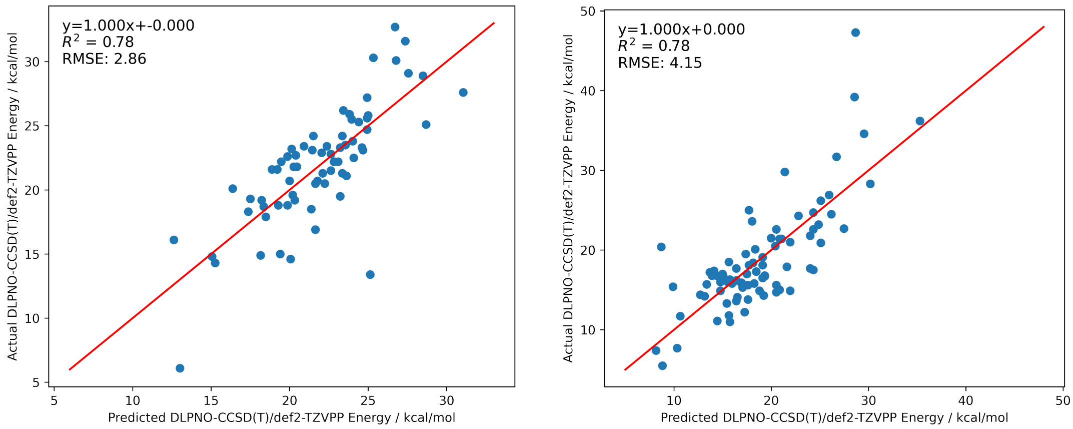


(a) Raw Values

(b) Scaled

Figure 3.12: Comparison of activation energies for B97-3c vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, including ligand exchange, TSOA and TSSig.

Figure 3.12 shows the correlation between B97-3c and DLPNO-CCSD(T)/def2-TZVPP relative energies. B97-3c energies correlate excellently with coupled cluster, with an R^2 value of 0.96. RMSE is superior to GFN2-xTB at 7.76 kcal mol^{-1} improving to 5.34 kcal mol^{-1} when scaling the energies using the equation of the line. B97-3c is much better than GFN2-xTB for calculating the Gibbs free energy of complexes in the Ullmann-Goldberg reaction.



(a) TSOA **(b)** TSSig

Figure 3.13: Comparison of activation energies for B97-3c vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, transition states only.

When only transition states are compared B97-3c correlates reasonably well with the DLPNO-CCSD(T) calculated activation energies, with a mean average error of 3.9 kcal mol^{-1} across both transition states, with 89% of structures falling with <1.5× RMSE. B97-3c performed slightly better for **TSOA** than **TSSig** with an RMSE of 2.86 and 4.15 kcal mol^{-1} respectively. This difference in RMSE is likely due to the bonding involved in the **TSSig** transition state not being as well described by the functional. The **TSOA** transition state involves a more common oxidative addition step which is likely present in a larger quantity in the functional training data due to it being common across several different reaction types. Whereas the **TSSig** transition state contains a far less common C−N stretched bond which is likely poorly described by the functional. Structures containing oximes and O-Cu-O 5-membered ring motifs generally correlate poorly with a >1.5× RMSE between the two methods. However, only 8 ligands containing oximes have been reported for the Ullmann-Goldberg reaction (***ligands_lit_set***), and this was not deemed a significant problem for ligand exploration. DFT methods were not tested as they were considered too computationally expensive as seen in the method benchmarking (see **Section 3.2.1**). Thus, B97-3c represents a good balance between computational time and accuracy for the calculation of activation energies for the Ullmann-Goldberg reaction. Based on this benchmark all optimisations and frequency calculations use GFN2-xTB. All energy calculations use the B97-3c composite method.

3.2.7.2 Success Rates

To analyse the reliability of the workflow the success rates for the generation of the initial structures and the optimisation of intermediate and transition state structures was determined. **Table 3.9** shows the success rate for the generation of initial structures for all ligands in the ***ligands_lit_set*** dataset. All ligands were supplied as SMILES strings.

Table 3.9: Success rate for the generation of organometallic complexes for all ligands in the ***ligands_lit_set*** dataset.

Structure	Success Rate (%)
CuLpip	96
CuLI	96
TSOA	86
TSSig	93

The success rate for initial structure generation is generally good across all structures. **TSOA** shows a much lower success rate than the intermediate structures and **TSSig**. The majority of failed structures were structures containing monodentate ligands. This is likely due to the increased steric bulk around the metal centre for the **TSOA** transition state and increased steric demand of two monodentate ligands around the copper centre leading to failed ligand insertion or incorrect ligand structures. **Table 3.10** shows the success rates with bidentate ligands only.

Table 3.10: Success rate for the generation of organometallic complexes for only bidentate ligands in the ***ligands_lit_set*** dataset.

Structure	Success Rate (%)
CuLpip	98
CuLI	98
TSOA	97
TSSig	99

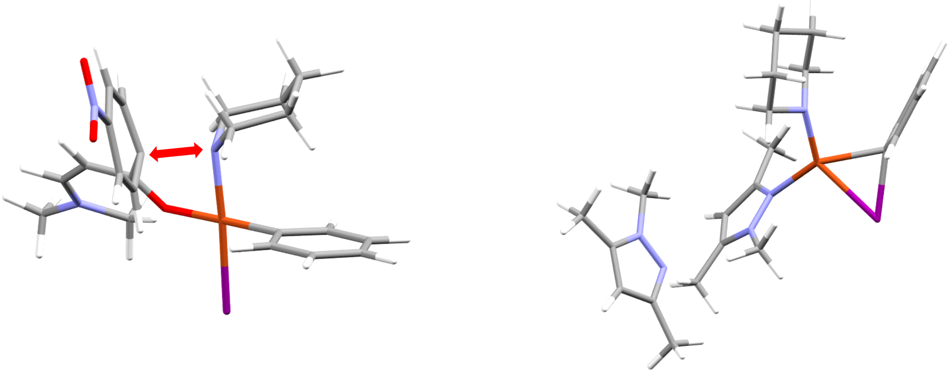
When bidentate ligands only are considered, the success rate of initial structure generation is much higher. The success rate of intermediate structures increases by 2% to 98%. A significant improvement is observed in both transition state starting structures with an improvement of 11% and 6% for **TSOA** and **TSSig** respectively. All success rates are above 97% which shows excellent reliability for the structure generation step. Due to the nature of the low stability of ligands in the Ullmann-Goldberg reaction only bidentate ligands are of interest, making the success rate excellent for this purpose.

Table 3.11: Success rate for the optimisation of organometallic complexes for only bidentate ligands in the ***ligands_lit_set*** dataset.

Structure	Success Rate (%)
CuLpip	93
CuLI	93
TSOA	88
TSSig	94

The second aspect of reliability in the workflow is the reliable optimisation of structures to either a minimum or a transition state. Intermediate structures require no imaginary frequencies and transition states require one imaginary frequency corresponding to the correct bond-forming process (Cu−C for **TSOA** and C−N for **TSSig**). A correct transition state was determined via manual inspection of the imaginary frequency. The success rates for each structure are summarised in **Table 3.11**. Successful structure optimisation was generally very good across all structures, both intermediates and transition states. The common cause of failure for intermediate structures **CuLpip** and **CuLI** was the presence of small imaginary frequencies $0 > x >$-40 cm^{-1}. To improve the reliability of these optimisations the convergence criteria were tightened to use the *TightOpt* criteria in ORCA. While iterative optimisation methods, where the structure is automatically distorted along the imaginary frequency and re-optimised, are available in ORCA they currently do not work with the extended tight-binding methods. When these methods become available, a further increase in optimisation reliability can be expected.

For transition states, **TSOA** shows a lower success rate than **TSSig**. This is likely due to the shallow potential energy surface in these transition states as observed previously in **Section 3.2.4**, making the location of the transition state difficult. In some cases, it may also be possible that the **TSOA** transition state does not exist for specific ligands as it follows a different mechanistic pathway. Overall the optimisation success rates are very good considering the number of calculations required.



(a) Reaction of the ligand **(b)** Ligand dissociation

Figure 3.14: Example incorrect transition states for the **TSOA** pathway.

Common causes of incorrect transition states were dissociation of the ligand (**Figure 3.14b**), the inability to find the transition state or incorrect transition states, such as a reaction between the ligand and the substrates (**Figure 3.14a**). Overall the success rates are very good for a high-throughput workflow and show excellent reliability across both intermediate and transition state structures.

3.2.7.3 Calculation Time

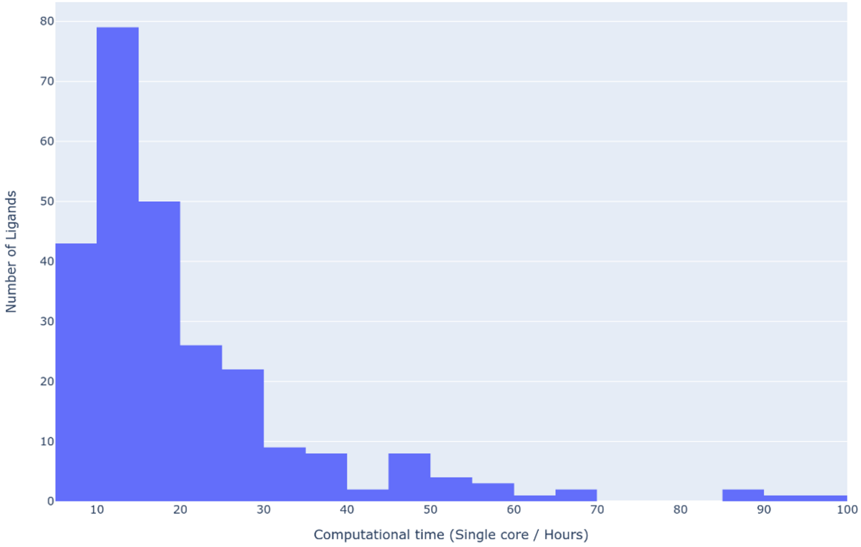


Figure 3.15: Histogram of the single core computational time required to calculate the activation energy for all ligands in the ***ligands_lit_set*** dataset.

The total single core computational time was calculated for each ligand in the ***ligands_lit_set*** to determine the average time taken per ligand. Each activation energy requires 4 optimisations, 2 transition state optimisations, 4 frequency and 4 energy calculations. Each calculation used 4 CPU cores and 4GB of RAM. The distribution of total calculation time is shown in **Figure 3.15**. The average time taken to calculate the activation energy for a single ligand is approximately 15h of single core computational time. For very large ligands, calculations can take >50 CPU hours. A similar full DFT calculation would take anywhere from a few days to a few months per ligand depending on the size of the ligand and the number of optimisation steps required, especially where a hessian calculation is required. This method provides a good balance between accuracy, reliability and both computational time and resources, making it ideal for a high-throughput predictive workflow.

3.2.7.4 Asymmetric Ligands

For asymmetric ligands, both isomers need to be considered when calculating the activity of the ligand. One orientation of the ligand in the complex may be significantly more active than the other, especially in some cases such as asymmetric catalysis. The need to calculate an inverted structure is dependent on the geometry of the complexes of interest. For example, three-coordinate trigonal planar complexes are mirror images upon ligand inversion and therefore do not need to be recalculated. Transition state structures, in the case of the Ullmann-Goldberg reaction, have two possible isomers which need to be compared to predict the activity of the ligand (**Figure 3.16**).

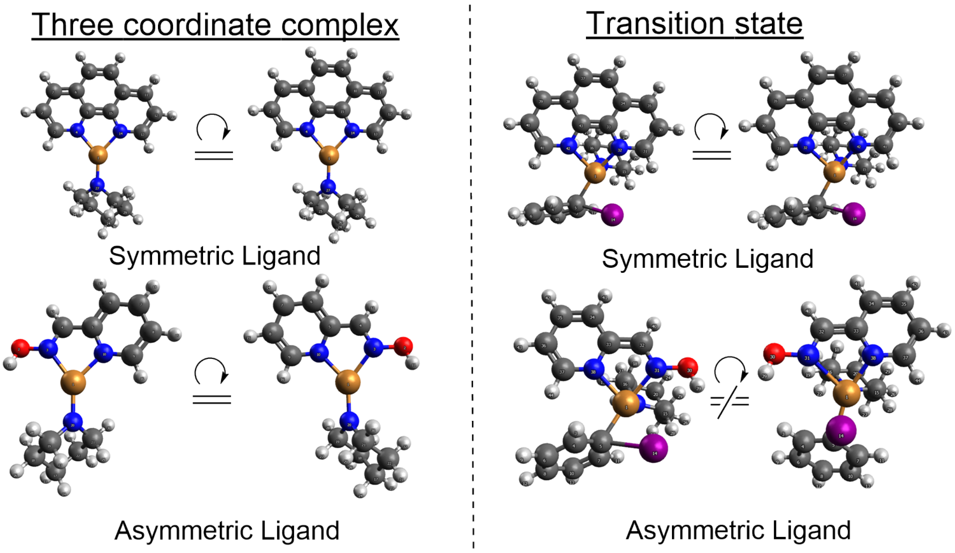


Figure 3.16: Comparison of three coordinate intermediate and transition state structures upon ligand inversion for both symmetric and asymmetric ligands.

In order to assess the need to generate both structures for each transition state in the computational workflow, all asymmetric ligands in the ***ligands_lit_set*** were used to compare the activation energies for both isomers. Both isomers were generated with the same structure generation procedure described in **Section 3.2.5** including the forced location of the ligand to ensure that the ligand was in the opposite orientation. The structures were optimised to a transition state using GFN2-xTB and their activation energies were calculated at the B97-3c level of theory. The comparison of activation energies of the original and inverted ligand is shown in **Figure 3.17**.

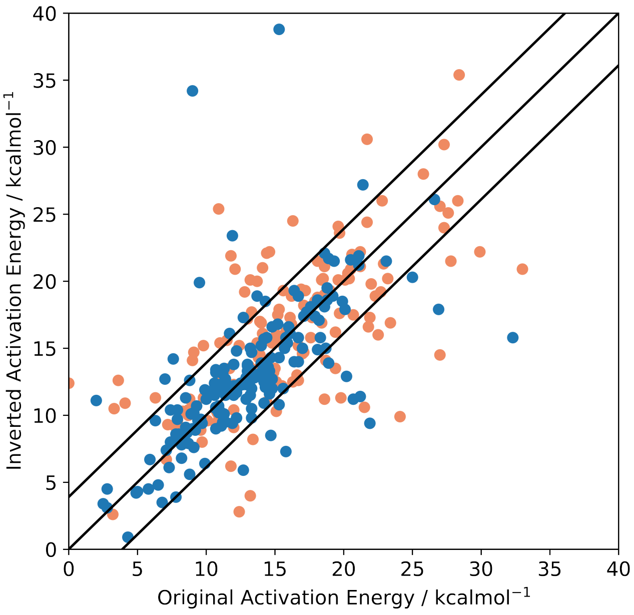


Figure 3.17: Activation Energy of the original **TSOA** (orange) and **TSSig** (blue) transition states compared to their activation energies upon ligand inversion for all asymmetric ligands in the ***ligands_lit_set***. Black lines show parity and ± 3.9 kcal mol^{-1}, the determined error in the calculation. Activation energies were calculated at the B97-3c level of theory.

The correlation of original activation energy and activation energy upon ligand inversion has an average difference of 2.3 kcal mol^{-1} (2.8 and 1.9 kcal mol^{-1} for **TSOA** and **TSSig** respectively) between isomers and is within the error of the calculation, 3.9 kcal mol^{-1} (**Section 3.2.7.1**). As the Ullmann-Goldberg reaction is a symmetric reaction, only one isomer needs to be generated. The need to generate only one isomer significantly reduces the computational time required. As the transition state calculation is the most computationally expensive step of the prediction workflow, doubling the number of this type of calculation would almost double the computational time and resources required per ligand.

For asymmetric synthesis tasks, however, it is recommended to generate both isomers.

3.2.8 Final Computational Workflow

The following section provides a complete step-by-step walkthrough of the full computational workflow. A summary of the workflow is shown in **Figure 3.18**. The workflow requires Python 3, ORCA version 4.2 or later, xtb version 6.3 or later and the custom version of molSimplify (see ESI).[3,93,167] Due to compatibility, the workflow requires a Linux-based operating system. This workflow assumes that the ligand structures and data file has been generated as described in **Chapter 2**. A *.csv* file containing ligand SMILES string and coordinating atom data may also be used. An example of this is provided in the ESI. All scripts used in this section are also provided in the ESI.

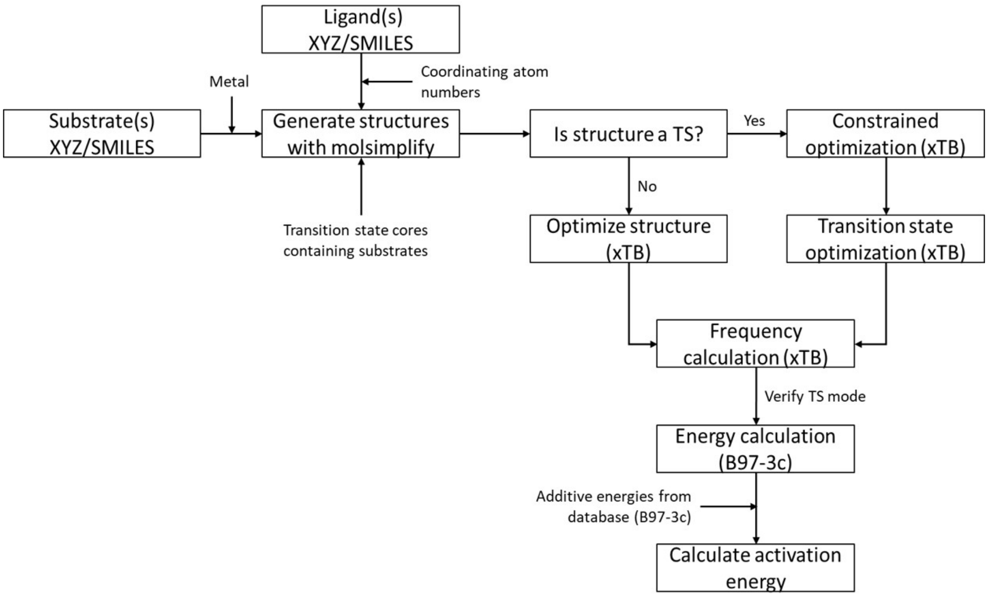

Ligand(s)
XYZ/SMILES
Coordinating atom numbers
Metal
Substrate(s)
XYZ/SMILES
Generate structures with molsimplify
Transition state cores containing substrates
Is structure a TS?
Yes
No
Constrained optimization (xTB)
Optimize structure (xTB)
Transition state optimization (xTB)
Frequency calculation (xTB)
Verify TS mode
Energy calculation (B97-3c)
Additive energies from database (B97-3c)
Calculate activation energy

Figure 3.18: Summary flow chart of the high-throughput computational workflow for the calculation of activation energies from ligand structure files.

3.2.8.1 Structure Generation

Before structure generation, all 3D structures files from the CSD-CrossMiner search should be copied to the `molSimplfy/Ligands/` folder. The contents of the *{search}.dict* file should also be copied into the *ligands.dict* file. Transition state cores also need to be generated at the desired level of theory and added to the `molSimplfy/Cores/` folder and *cores.dict* file along with the atom indexes of the atoms to be replaced. If custom deprotonation rules are being used the SMARTS strings need to be placed on line 538 of the `molSimplify`

/Classes/globalvars.py file in the molSimplify installation folder. MolSimplify input files are generated using the TS_mols_gen_xyz.py script. This script contains all of the information required to generate the correct organometallic structures using molSimplify. If the script is being used for another reaction this file needs to be edited to generate the correct structures as described in **Section 3.2.5**. MolSimplfy input files can then be generated using the following command:

```
python TS_mols_gen_xyz.py {search}.csv
```

where {search}.csv is the file containing the ligand data obtained from the CSD-CrossMiner search. Structures can then be generated by running the following bash script in the folder containing all of the molSimplify input files.

```
for f in *.inp; do
    molsimplify -i {f};
done
```

3.2.8.2 Generating Computational Input files and Running Calculations

The input files for the computational calculations are generated automatically using the TS_input_gen_HPC_constrain.py script. This script contains all of the parameters used in the computational calculations such as methods, solvent, computational resource requirements and atom constraints for the TS optimisation. If another reaction is being studied this script needs to be edited to provide the correct parameters and TS active atom constraints. Charge, spin and 3D coordinates are automatically identified from the molSimplify output files. By default, all optimisation and frequency calculations use the GFN2-xTB method and all energy calculations use the B97-3c method. Other computational methods can be used if required. Computational input files can then be generated by running the following bash script:

```
cd molsimplify_input_files || exit

cd Runs || exit

for dir in */; do
    if [ -d "$dir" ]; then
        cd "$dir" || exit
        for dir2 in */; do
            if [ -d "$dir2" ]; then
                cd "$dir2" || exit
                python ../../../../TS_input_gen_HPC_constrain.py
                cd ..
            fi
        done
        cd ..
    fi
done
```

Computational calculations can then be run by either looping through the Runs/ folder or by copying the folder to a high-performance computing facility and running the calculations through a batch queuing system. Calculations should be run sequentially in the order pre-optimisation > optimisation > frequency > energy. This is to prevent calculations from failing due to the previous calculation not being complete.

Once the calculations have finished the Runs/ folder can then be cleaned using the provided script to remove any unneeded output files. This drastically reduces the size of the folder allowing for easier copying if using a high-performance computing facility. Energies for additional structures such as additives, products and reactants need to be calculated separately.

3.2.8.3 Structure Verification and Calculation of Activation Energies

To verify if the optimised structures are correct the ts_check.py script is used by looping through the folders using the same bash script as above but replacing

TS_input_gen_HPC_contrain.py with ts_check.py. The values of S_0 and tolerances for the magnitude of imaginary frequencies can be adjusted. If another reaction is being studied, the atom indexes of the TS active atoms need to be changed to those in the TS core of interest.

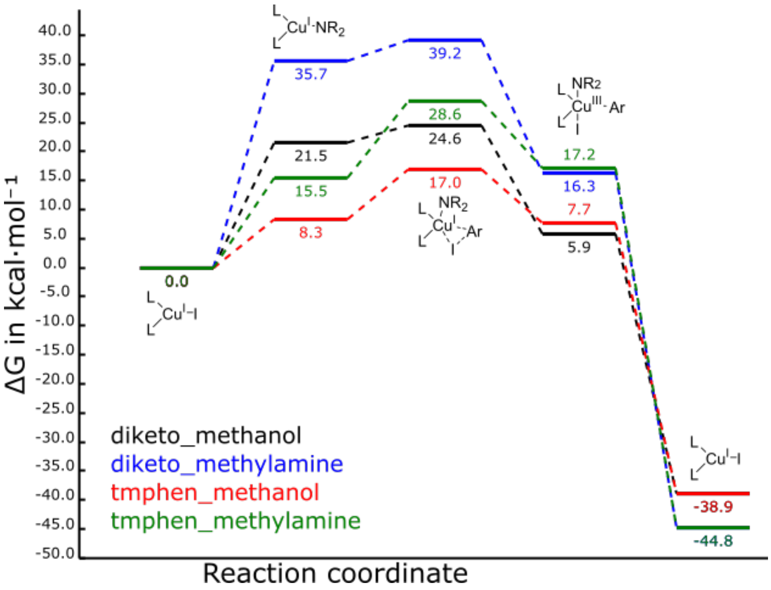


Figure 3.19: Example energy profiles for the oxidative addition transition state for two ligands from the Buchwald et al. (2010) study with methylamine and methanol nucleophiles.[85]

Activation energies can then be calculated by running the Energy_Analysis_CrossMiner .py script. Again, if another reaction is being studied this script needs to be adjusted to ensure the calculation of the activation energy is correct. Activation energies are calculated and output in a *.csv* file. **Figure 3.19** shows an example output from the workflow for the Ullmann-Goldberg reaction using the Buchwald et al. ligands as an energy profile for the oxidative addition pathway.[85]

3.3 Conclusions

The high-throughput computation prediction of ligand activity for organometallic catalysts is currently limited by the lack of accurate low-cost computational methods as well as the lack of reliable methods for generating and optimising transition states. The development of accurate, reliable and low-cost tools will aid the discovery of new organometallic catalysts computationally.

This chapter presented a novel semi-automated computational workflow for the prediction of catalyst activity which (i) automatically generates both intermediate and transition state organometallic complexes, (ii) performs computational calculations for the determination of structural and electronic properties, (iii) analyses and predicts the activation energy for each ligand. The workflow is flexible allowing it to be used for any reaction of interest by modification of the source code.

The Ullmann-Goldberg reaction was used to develop and test the workflow. The activation energies for 345 ligands were calculated. The resulting activation energies were benchmarked against the 'gold-standard' coupled-cluster method, with an error of 3.9 kcal mol^{-1} at a much lower computational cost than traditional DFT.

This workflow offers several advantages over currently used methods due to its faster speed and lower computational cost, coupled with good accuracy compared to higher-level methods. This workflow has wide applicability in catalyst design, ranging from pharmaceutical process development, mechanistic exploration and novel catalyst design. It can also be applied in various chemical areas such as pharmaceuticals, agrochemicals or chemical discovery.

3.4 Methodology

All scripts used in this chapter are available on GitHub at `https://github.com/MarcS18/Thesis_ESI`.

3.4.1 Computational Methods

Semi-empirical and DFT methods were performed using Gaussian09.[168] Composite and coupled cluster methods were performed using ORCA 4.2.1.[3] The University of Leeds supercomputer, ARC3, was used on standard nodes with 24-core Broadwell E5-2650v4 CPUs at 2.2 GHz with turbo and 128 GB of memory.[169] Complexes were optimised in the gas phase using the default convergence criteria unless stated otherwise. Extended Tight Binding calculations were performed with xtb 6.3.2[167] on an AMD Ryzen 3900X 12-core CPU at 3.8 GHz with turbo and 16 GB of memory. XTB calculations were optimised to the *tight* criteria unless stated otherwise. The solvent model based on density (SMD) implicit solvent model was used for B97-3c and DLPNO-CCSD(T) and the generalised Born model with surface area contributions (GBSA) implicit solvation model was used for GFN2-xTB. DMF was used as the solvent in all cases.

3.A Appendix

3.A.1 molSimplify

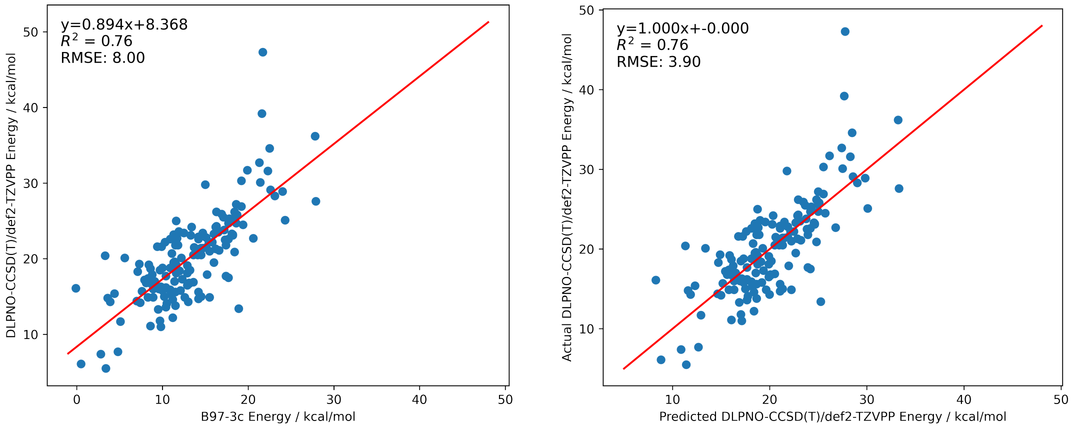
3.A.1.1 Geometry Values

Possible values for `-coord`: 1, 2, 3, 4, 5, 6, 7, 8.

Table 3.A.1: Possible values for `-geometry`.

Value	Geometry
no	None
li	Linear
tpl	Trigonal Planar
sqp	Square Planar
thd	Tetrahedral
spy	Square Pyramidal
tbp	Trigonal Bi-pyramidal
oct	Octahedral
tpr	Trigonal Prismatic
pbp	Pentagonal Bipyramidal
sqap	Square Antiprismatic

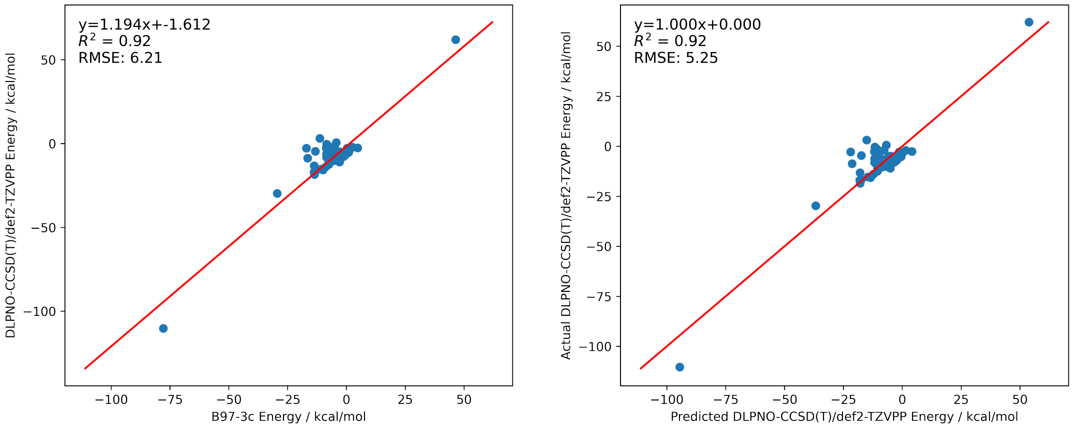
3.A.2 Activation Energy Benchmarking



(a) Raw Values

(b) Scaled

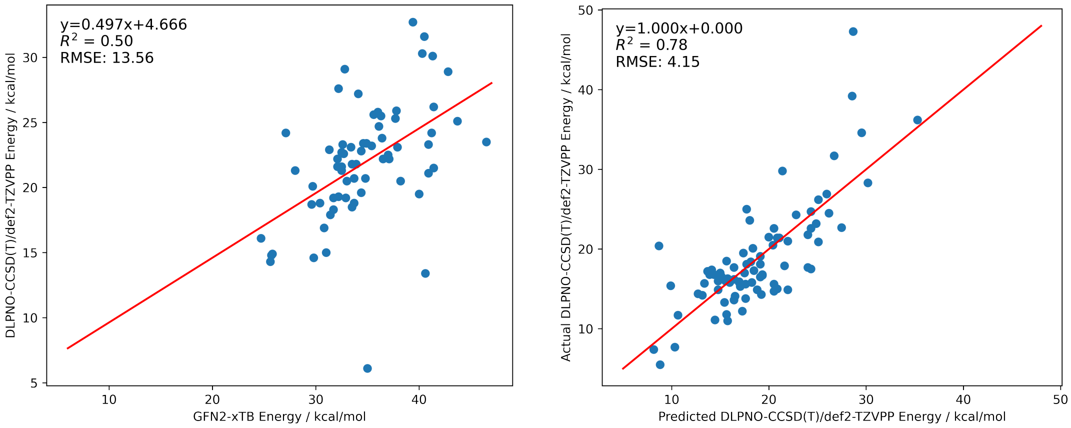
Figure 3.A.1: Comparison of activation energies for B97-3c vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, combined transition states.



(a) Raw Values

(b) TSSig

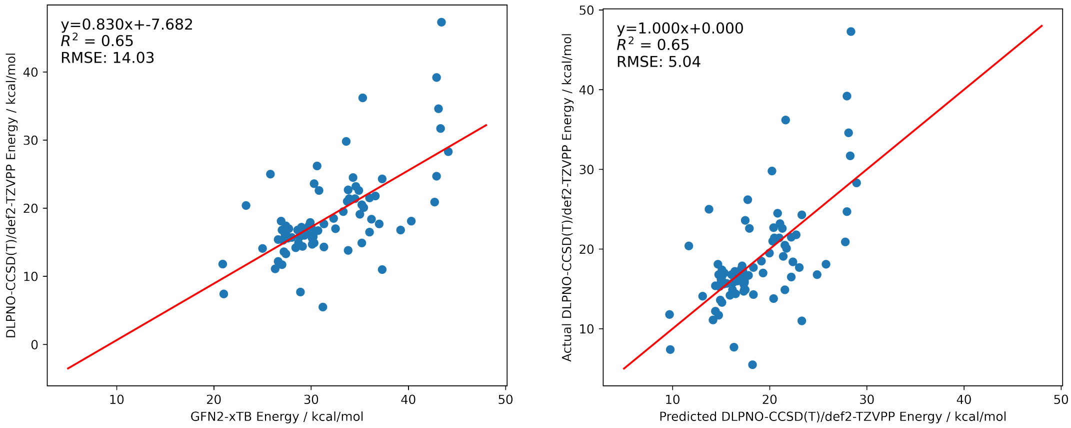
Figure 3.A.2: Comparison of activation energies for B97-3c vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, ligand exchange only.



(a) Raw Values

(b) Scaled

Figure 3.A.3: Comparison of activation energies for GFN2-xTB vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, TSOA only.



(a) Raw Values

(b) Scaled

Figure 3.A.4: Comparison of activation energies for GFN2-xTB vs DLPNO-CCSD(T)/def2-TZVPP for 100 literature ligands, TSSig only.

Chapter 4: Application of CatSD to the Ullmann-Goldberg Reaction

4.1 Introduction

4.1.1 Reaction Mechanisms

The first mechanistic hypotheses for the Ullmann-Goldberg reaction began to emerge in the 1960s, with several different mechanisms being proposed until the 1990s.[72,75,170–173] While it was generally agreed that the coordination of the nucleophile to the copper was involved, the process for the activation of the aryl halide was under dispute. The observed reactivity of the aryl halide followed the order I > Br > Cl, which is the opposite of the observed reactivity for common aromatic nucleophilic substitution reactions. Therefore, it was obvious that the metal was involved in some way in activating the aryl halide.

The proposed mechanisms can be divided into four main classes:

1. Radical mechanisms.
2. σ-metathesis mechanisms.
3. π-complexation mechanisms.
4. Oxidative addition/reductive elimination mechanisms.

4.1.1.1 Radical based mechanisms

The first postulation of free-radical involvement in the Ullmann-Goldberg reaction was by Waters in 1937.[174] However, it was not until 1970 that the first radical mechanism was proposed by Bunnett.[170] He proposed a radical type aromatic nucleophilic substitution ($S_{RN}1$)

after studying the reaction of iodoarenes with potassium amide. Bunnett's mechanism is initiated by the single electron transfer from the 'outer sphere', where the initiator does not coordinate to the aryl halide, to form a radical anion (**Scheme 9**). While Bunnett's study was not conducted on a system containing a metal catalyst, any metal that can undergo single electron transfer would be suitable as an initiator. Copper can both act as a single electron oxidant in many reactions such as dehydrogenative functionalisation, as well as a single electron reductant.[175,176]

$$(\mathrm{ArX})^{\bullet-} \longrightarrow \mathrm{Ar}^{\bullet} + \mathrm{X}^{-} \tag{4.1}$$

$$\mathrm{Ar}^{\bullet} + \mathrm{Nu}^{-} \longrightarrow (\mathrm{ArNu})^{\bullet-} \tag{4.2}$$

$$(\mathrm{ArNu})^{\bullet-} \longrightarrow \mathrm{ArNu} + (\mathrm{ArX})^{\bullet-} \tag{4.3}$$

Scheme 9: General reaction scheme for the $S_{RN}1$ mechanism proposed by Bunnet.[170]

A few years later, another radical mechanism was proposed by Kochi and Jenkins, called Halogen Atom Electron Transfer (HAT) or 'inner sphere' electron transfer.[171] In comparison to Bunnett's mechanism the aryl radical is formed by the transfer of a neutral halogen atom from the aryl halide to the copper atom (**Scheme 10**).

$$\mathrm{Cu^{I}I} + \mathrm{Nu}^{-} \longrightarrow \mathrm{Cu^{I}Nu} + \mathrm{X}^{-} \tag{4.4}$$

$$\mathrm{ArX} + \mathrm{Cu^{I}Nu} \longrightarrow \mathrm{Ar}^{\bullet} + \mathrm{Cu^{II}XNu} \tag{4.5}$$

$$\mathrm{Ar}^{\bullet} + \mathrm{Cu^{II}XNu} \longrightarrow \mathrm{ArNu} + \mathrm{Cu^{I}X} \tag{4.6}$$

Scheme 10: General reaction scheme for Halogen Atom Transfer via the $S_{RN}1$ mechanism.

The first evidence of a radical mechanism emerged in 1978 from a study by Arai et al. who were studying the reaction of 1-bromoanthraquinone with 2-aminoethanol, catalysed by copper bromide. Formation of a paramagnetic species was detected by EPR experiments, which was identified as the 1-bromoanthraquinone radical anion, formed by electron transfer from the Cu(I) species (**Scheme 11**). The production of anthraquinone was also detected and was explained by the dehalogenation of the 1-bromoanthraquinone radical anion. This was the first study that supported the $S_{RN}1$ radical pathway.

However, several authors published evidence against the formation of radicals during the reaction. A comparative study between different $S_{RN}1$ processes by Bowman found that Cu-catalysed coupling was the most efficient for the synthesis of heterocycles, however,

$$\mathrm{Cu^{I}X + ArX \longrightarrow Cu^{II}X_2 + Ar^{\bullet}} \tag{4.7}$$

$$\mathrm{Ar^{\bullet} + Nu^{-} \longrightarrow (ArNu)^{\bullet -}} \tag{4.8}$$

$$\mathrm{(ArNu)^{\bullet -} + Cu^{II}X_2 \longrightarrow ArNu + Cu^{I} + X^{-}} \tag{4.9}$$

Scheme 11: General reaction scheme for copper catalysed $\mathrm{S_{RN}1}$ mechanism.

showed many differences from radical processes.[177,178] Firstly, the addition of radical scavengers or oxygen did not inhibit the reaction and therefore, led the authors to believe that a radical process was not involved.[177,178] Secondly, multiple radical clock experiments have been used to investigate the existence of the aryl radical intermediate, all with negative results, except for one experiment by Fier and Hartwig.[179] Fier and Hartwig observed that a small amount of radical-derived product was formed, however only in small quantities, showing that radical mechanisms can occur, but do not govern the reaction mechanism.[179]

Radical clock experiments are based upon the fact that under a radical mechanism, the Ullmann-Goldberg coupling of the nucleophile should be much faster than the most kinetically favoured 5-*exo*-trig ring closure of a methylcyclopentane moiety.[180] This has been proposed several times but has yet to be demonstrated.[181]

Another set of experiments by Hartwig et al. used a set of aryl chlorides and bromides with higher reduction potentials than the aryl iodide.[182] Therefore, if a radical mechanism was involved the reaction rate should increase with increasing reduction potential but it was observed that the less reducible aryl iodides had higher reaction rates, leading the authors to exclude a radical mechanism.[182]

In 2010, Buchwald contested both of these experimental methodologies. He suggested that the intermediate aryl radical could not exist as a free radical but instead existed as a caged radical pair and therefore, is unable to react with the alkene in the radical clock experiments, invalidating the results of these tests.[85] He also explained Hartwig's experiments in terms of the less effective coordination of these substrates to copper.[85]

4.1.1.2 σ-bond metathesis

Bacon and Hill proposed in the mid-1960's a mechanism in which the copper forms a four-centred σ-complex with a lone pair of electrons on the halogen atom inducing polarisation of the C−X bond, facilitating the attack of the nucleophile.[173] However, such a mechanism is not easy to prove as multi-centre processes are hard to differentiate from those involving ionic intermediates. Litvak and Shein put forward an adaption of this mechanism whereby they combined the four-centre intermediate with a radical process, and later Aalten et al. reported a similar mechanism that proceeds through an intimate-electron transfer, using the Cu(I)/Cu(II) redox couple.[75,172]

$$\mathrm{ArBr} + \mathrm{Cu^{I}(OR)L_{n}} \underset{+\mathrm{L}}{\overset{-\mathrm{L}}{\rightleftharpoons}} \mathrm{ArBrCu^{I}(OR)L_{n-1}} \rightleftharpoons \mathrm{ArBr^{\bullet -}Cu^{II}(OR)L_{n-1}} \rightleftharpoons \tag{4.10}$$

$$\mathrm{ArBr^{\bullet -}Cu^{I}(OR^{\bullet})L_{n-1}} \rightleftharpoons \mathrm{Ar}^{\bullet}\underset{\mathrm{OR}^{\bullet}}{\overset{\mathrm{Br}^{-}}{\diamond}}\mathrm{Cu^{I}L_{n-1}} \rightleftharpoons \mathrm{ArOR} + \mathrm{Cu^{I}BrL_{n-1}} \tag{4.11}$$

Scheme 12: Litvak's proposed mechanism for the copper-catalysed etherification reaction.[172]

In general, this mechanistic pathway can be summarised by **Scheme 12**. The first step of the displacement of the halide by the nucleophile to form the catalytic species. The copper then coordinates to the aryl halide via a four-centred intermediate. Coordination generates a partial positive charge on the C−X carbon through polarization of the C−X bond, assisting the substitution by the nucleophile. Copper remains in the +1 oxidation state throughout the entire catalytic cycle. In 2004, a non-radical based σ-bond metathesis pathway was proposed by Van Allen.[183]

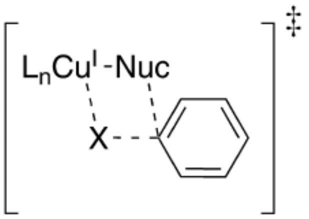
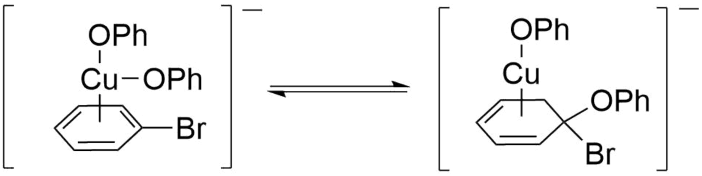


Figure 4.1: Van Allen's proposed 4-centre sigma metathesis transition state.[85]

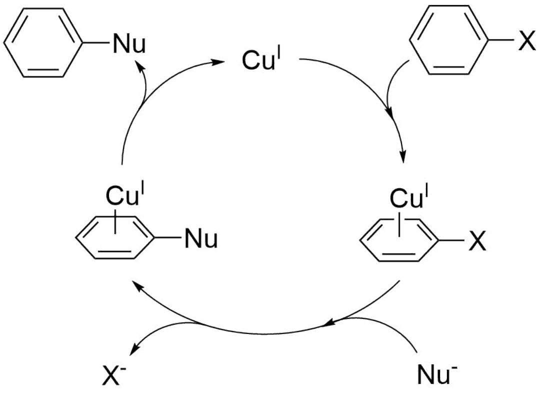
4.1.1.3 π-complexation

In Weingarten's work reporting the activity of Cu(I), he also proposed a mechanism for the activation of the aryl halide.[72] During the investigation of the reaction to form a phenyl ether from bromobenzene and potassium phenoxide, the existence of the $[\mathrm{Cu(OPh)_2}]^{-}$ was proposed. This copper species would coordinate to the π system of the aryl halide, essentially acting as an activating group, making the C−X bond more susceptible to nucleophilic substitution (**Scheme 13**). However, there is scant experimental evidence to support this hypothesis.



Scheme 13: Weingarten's proposal for the intermediate via π-complexation.

This pathway was deemed plausible for several reasons. First of all η^2-Cu(I)-benzene complexes had been synthesised the year before.[184] Secondly, η^6- haloarene-Cr(O) complexes had been shown to be very effective in aromatic nucleophilic substitution reactions.[185] Finally, calculations have shown that η^6 coordination is preferred over η^2 or η^1 coordination in complexes of copper and benzene.[186] However, in practice η^6 complexes are very rare, whereas η^2 complexes are more common between copper and aromatic ligands.[184,187] Moreover, Paine pointed out that this mechanism did not explain the accelerating effect observed with an *ortho* carboxylate group, whereas moving the carboxylate group to the *para* position gives no increase in reaction rate.[188] Also, the comparison of copper with Cr(0) complexes is not valid as the order of halide reactivity is reversed in the Cr(0) case with F and Cl being the most reactive aryl halides.[189] This mechanistic pathway can be summarised in the following catalytic cycle (**Scheme 14**):



Scheme 14: Proposed catalytic cycle involving the π-complex intermediate.

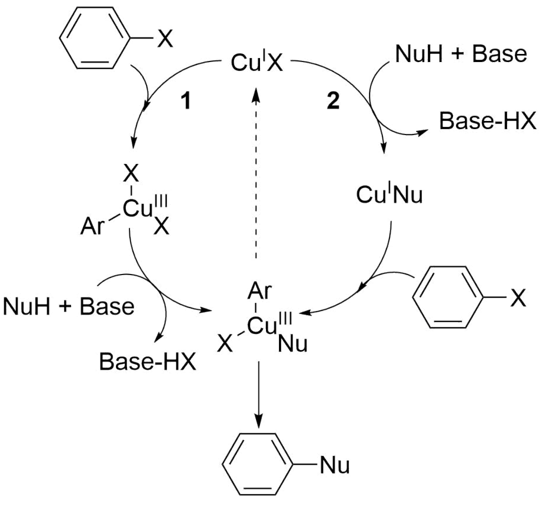
Cu(I) coordinates to the π system of the aromatic ring. The aryl halide then undergoes nucleophilic substitution due to polarization of the C−X bond. The coupling product then dissociated to restore the Cu(I) catalyst. Throughout the entire catalytic cycle, the copper remains in the +1 oxidation state.

4.1.1.4 Oxidative/reductive mechanisms

Several literature sources elicit the formation of Cu(III) intermediates in the Ullmann reaction. Cohen first proposed the existence of a Cu(III) intermediate, when studying the reaction of *o*-iodo-*N*,*N*,dimethylbenzamide with copper chloride and benzoic acid in DMF.[190] It was observed that increasing the concentration of benzoic acid increased the formation of *N*,*N*-dimethylbenzamide while decreasing the formation of the Cl-substituted *o*-chloro-*N*,*N*-dimethylbenzamide. It was also observed that upon the addition of copper chloride, the reverse is true. To explain this behaviour he proposed a mechanism via the oxidative addition of the aryl halide to the copper salt, leading to a Cu(III) intermediate.

Along with Cohen's proposed mechanism, he also ruled out the possibility of two other proposed mechanisms. The observed reactivities could not be explained by the formation of four-centred intermediates. A radical mechanism was also excluded as no *N*-methylbezamide was observed in the reaction, which would be formed from rapid hydrogen abstraction from the methyl group *ortho* to the aryl radical.

Following Cohen, Bethell investigated the reaction of primary amines with 1-halogenoanthraquinones promoted by copper salts.[191] Two products were detected, the aminated anthraquinone and the dehalogenated anthraquinone. It was observed that the reaction rate was dependent on the halogen (I > Br > Cl), but this had no effect on the ratio of the two products. Secondly, deuteration experiments showed that N-deuteration of the amine gave only a small kinetic isotope effect and did not affect the product ratio, whereas deuteration on the α carbon of the amine gave a large kinetic isotope effect leading to a large increase in the formation of the amine product. Finally, Bethell observed that the ratio of aminated to dehalogenated product was directly proportional to the concentration of the amine. Bethell explained these results by suggesting an intermediate arylcopper(III) complex with one or more amine ligands and one amide ligand. The ratio of aminated to dehalogenated products is determined by the competition between intramolecular hydrogen transfer from the C−H bond of the amide ligand and intermolecular amination.



Scheme 15: Two proposed pathways for the order of the oxidative addition and transmetallation steps.

There are currently two potential oxidative addition/reductive elimination pathways (**Scheme 15**). The first of which involves the oxidative addition of the aryl halide to the copper to form a copper(III) intermediate. The halide is exchanged with the nucleophile and reductive elimination releases the coupled product. The second and most favoured pathway is where the nucleophile reacts with the copper(I) halide before the oxidative addition takes place. Unlike palladium cross-couplings, where the oxidative addition step precedes transmetallation, in the copper system, the order of these steps is still under debate and either of the two routes could take place.[192]

The mechanism for the Ullmann condensation reaction is still uncertain and may be dependent on ligand properties, substrates and reaction conditions. The current generally accepted mechanism is that of the oxidative addition/reductive elimination pathway after coordination of the nucleophile as Cu(III) intermediates are not very stable and the oxidative addition step is promoted by the high electron density on the Cu atom in the Cu(I) intermediate.[193]

4.1.2 Modelling Transition States

4.1.2.1 Oxidation state of Copper

Three oxidation states of copper have been found to be effective in copper-catalysed cross-coupling reactions. Copper sources of oxidation states Cu(0), Cu(I) and Cu(II) have been successfully used in Ullmann-Goldberg reactions, with salts and oxides working well for several nucleophiles.[69] This observation suggests that a common copper species is formed during the reaction, and a lot of work in the early 1960s was aimed at the electrochemical behaviour of copper sources.[72,75,188] Weingarten in 1964 proposed that Cu(I) could be the common intermediate as it was observed that Cu(I) sources led to slightly higher reaction rates.[72] It was later demonstrated that Cu(II) species could be reduced to Cu(I) in the presence of coordinating molecules, with common nucleophiles such as phenoxides and amines acting as the redox counter partner.[72,188]

In 1987, Paine, using electron microscopy and X-ray powder diffraction, found that Cu(0) particles were covered in a layer of Cu_2O. He proposed that if this Cu_2O layer leached into the solution it would provide the required Cu(I) species for catalysis. Observation of the surface of the solid catalyst recovered after the reaction showed crystals of Cu_2O, which were not present on the original catalyst.[188] This supported the hypothesis of leaching, as the crystallisation of Cu_2O would only be possible if it had leached into the solution and then recrystallised after the reaction.[188]

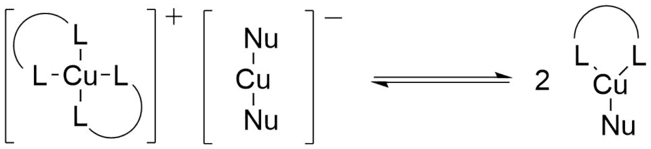
Evidence for the oxidation of Cu(0) to Cu(I) in the presence of a ligand such as phenanthroline was only recently observed by Taillefer et al. using in situ cyclic voltammetry.[194] Wei et al. also proposed that oxidation from Cu(0) to Cu(I) was possible by reaction with atmospheric oxygen, on the basis of colour change and catalytic results under different conditions for the reaction of aryl halides and amines in water using metallic copper powder as the copper source.[195]

These investigations demonstrate that Cu(I) is the active catalytic species however, the initial source of copper is suggested to be unimportant for the outcome of the reaction, due to the generation of Cu(I) in situ, via oxidation and reduction processes.[75]

4.1.2.2 Catalyst resting state

Identification of key reaction intermediates and, in particular, the catalyst resting state has been the focus of numerous investigations. Prior to the use of ancillary ligands in the reaction, stoichiometric amounts of a copper(I) source were used to couple heteroatom nucleophiles with organic halides to afford the C-heteroatom coupled product. Several groups have studied complexes of copper(I) alkoxides, phenoxides and amidate and their reactivity with alkyl and aryl halides to afford C−O and C−N coupled products. One of the first reports by Bacon and Karim studied the reaction of aryl halides with phthalimides.[196] It was observed that the most reactive system used a Cu:phthalimide ratio of less than one, where increasing the quantity of phthalimide reduced reactivity. This was the first example supporting a mechanism where the active species is a copper(I) mononucleophile complex, with the copper(I) dinucleophile complex being unreactive. Further evidence supporting the copper(I) mononucleophile intermediate species was provided by Paine when studying the reaction of diphenylamine with aryl halides to form triphenylamines.[188] Paine observed that the reaction was zero order with respect to the amine nucleophile, suggesting that the copper(I) species in solution coordinates to the nucleophile in a fast and irreversible step to give the intermediate species [Ph_2NCu^I], which reacts with iodobenzene in the rate-determining step. This observation along with the work of Bacon and Karim suggests that the catalyst resting state is a copper(I) species ligated by one deprotonated nucleophile ligand.

A more recent study by Buchwald and co-workers expanded on the work from Bacon and Karim, investigating the effect of ligand concentration on the active copper species in solution.[81] They studied the role of 1,2-diamine ligands on the *N*-arylation of amides. When a low concentration of diamine ligand is used ($[L] = 0.04$ M) an inverse rate dependence of amide concentration is observed due to the preferential formation of the unreactive diamidate copper complex over the active diamine copper(I) amidate. High concentrations of diamine ligand ($[L] = 0.28$ M) gave the diamine copper(I) halide complex as the major species in solution, which readily converts to the active amidate complex. To confirm the proposed catalyst resting state copper(I) pyrrolidinoate was reacted stoichiometrically with 3,5-dimethyliodobenzene in the presence of the diamine ligand yielding the coupling product.[82] This was further confirmed by DFT calculation by Guo and co-workers, validating the reactivity of the [LCu(NHAc)] complex.[197] Jutand and co-workers identified the same structure for the catalyst resting state as proposed by Buchwald, with a 1:1:1 ratio of ligand, copper(I) and deprotonated nucleophile when studying the reaction of iodobenzene with cyclohexylamine by NMR, electrochemistry and DFT, in the presence of a 1,3-diketone ligand.[198–200] Other publications by Hartwig, Peters and Fu also determined that the catalyst resting state had the same trigonal planar geometry.[181,182,201]

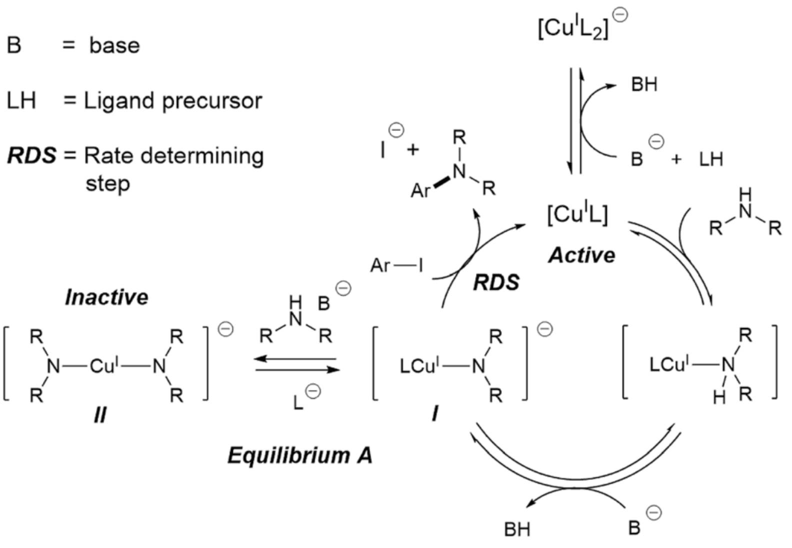


Scheme 16: Equilibrium between dimeric [L_2Cu][Cu(Nu)$_2$] and neutral monomeric [LCu(Nu)].

To determine the species present in solution Hartwig and co-workers characterized several copper(I) imidate and amidate complexes with bidentate N,N and P,P auxiliary ligands.[182] The species present in solution was determined to be a dimeric [L_2Cu][Cu(Nu)$_2$] ionic species in equilibrium with the neutral monomeric [LCu(Nu)] species, as determined by NMR spectroscopy and conductivity experiments. All tested complexes successfully reacted with iodobenzene to give their C−N coupled products in high yields, while the anionic species [Cu(Nu)$_2$]$^-$ present in the form [Cu(Nu)$_2$][Bu_4N] was unreactive. Therefore, the three-coordinate [LCuNu] complex was proposed to be the active intermediate. Similar studies with copper(I) phenoxide complexes with phenanthroline and cyclohexadiamine auxiliary ligands gave the same conclusion of a three-coordinate active intermediate.[201]

Expanding on bidentate ligands Taillefer and co-workers used a tetradentate bis(iminopyridine) ligand.[202] The coordination of the tetradentate ligand to copper(I) iodide in acetonitrile leads to the formation of a highly insoluble dimeric complex ([Cu_2L_2]I_2). The small fraction of dimeric complex available in solution forms the active monomeric copper(I) precatalyst upon displacement. This was evidence that the ligand does not necessarily need to aid the solubilisation of the copper(I) complex. The formation of insoluble dimeric species creates a reservoir of copper(I) protected from degradation processes, which would otherwise occur in solution.

The majority of the literature favours a trigonal planar complex where the deprotonated nucleophile is coordinated to the Cu(I) centre before activation of the aryl halides occurs. The addition of auxiliary ligands prevents the formation of less reactive, multiply-ligated copper complexes, ensuring the active catalyst resting state is present in high concentrations.[81,82,197]



Scheme 17: Generally accepted Ullmann-Goldberg reaction mechanism.[203]

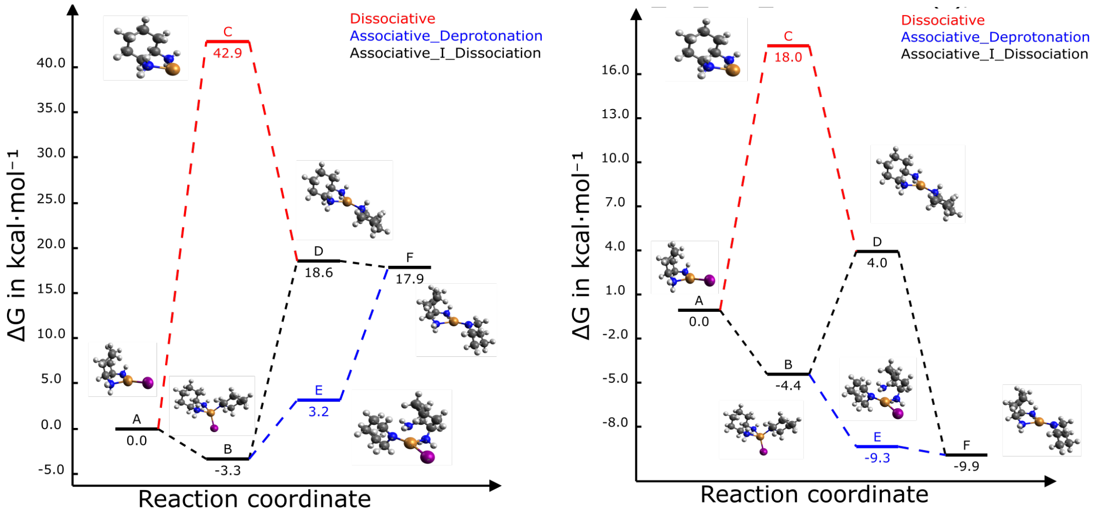
4.2 Results and Discussion

4.2.1 Ligand Exchange Mechanism

Identification of the rate-determining step is important for calculating the correct activation energy for a reaction. Calculation of the wrong step in the mechanistic pathway doesn't represent the energy barrier required for the reaction to proceed. If there is a reaction intermediate whose energy is lower than the initial reactants, then the activation energy needed to pass through any subsequent transition state depends on the Gibbs free energy of that transition state relative to the lower-energy intermediate. The rate-determining step is then the step with the largest Gibbs energy difference relative either to the starting material or to any previous intermediate on the diagram.

For the Ullmann-Goldberg reaction, the sigma metathesis pathway contains only one transition state, which is, therefore, likely the rate-determining step. For the oxidative addition/reductive elimination pathway, it has been shown previously that the oxidative addition step is the rate-determining step.[85] To confirm the rate-determining step for both pathways the ligand exchange step was modelled to verify that the exchange between the iodide and nucleophile to form the active catalytic species is not the rate-determining step. Three exchange mechanisms were modelled: dissociative (iodide dissociated, followed by coordination of the nucleophile then deprotonation), associative deprotonation (coordination of the nucleophile followed by deprotonation then dissociation of the iodide) and associative I dissociation (coordination of the nucleophile followed by dissociation of the iodide then deprotonation of the nucleophile).

All three ligand exchange pathways were calculated for 10 ligands from ***ligands_lit_set***. Ligands consisted of a range of N, O- donor atoms and functional groups as well as ligand charges (-2 to +1). Structures were optimised at the GFN2-xTB level of theory and energies were calculated at the B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory. The energy diagrams for ligand **L0003** are shown in **Figure 4.2**. Energy diagrams for the remaining ligands can be found in **Appendix 4.A.1**.



(a) GFN2-xTB//B97-3c **(b)** DLPNO-CCSD(T)/def2-TZVPP

Figure 4.2: Ligand exchange mechanisms for ligand **L0003** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and GFN2-xTB//DLPNO-CCSD(T)/def2-TZVPP levels of theory.

The associative deprotonation pathway is the lowest energy pathway for 9 out of the 10 ligands. For the remaining ligand, the associative I dissociation pathway is the lowest in energy. This ligand contains a BINOL structure and is the only ligand with a -2 charge, therefore, ligands with a -2 charge likely proceed with dissociation of the iodide before deprotonation. This is likely due to the large build-up of negative charge on the copper centre if the nucleophile is deprotonated before the iodide dissociates, resulting in a net -3 charge (Cu(I)). For ligands with a charge of +1, 0 and -1, deprotonation of the nucleophiles is favoured before iodide dissociation. This is likely due to coordination promoting the deprotonation of the nucleophile and the ability of copper the stabilise the additional negative charge on the nitrogen.

In all cases, ligand exchange does not proceed via a high-energy transition state. The process is either energetically favourable ($\Delta G < 0$ kcal mol^{-1}) or proceeds through a low energy intermediate ($\Delta G < 10$ kcal mol^{-1}). This observation is consistent across both the B97-3c and DLPNO-CCSD(T) calculations. This verifies that the ligand exchange between iodide and the nucleophile to generate the active catalytic species is not the rate-determining step of the reaction and therefore, does not need to be modelled to predict activity.

4.2.2 Analysis of Literature Ligands

The literature ligand set (***ligands_lit_set***) was analysed to identify the most common structural motifs in frequently used ligands in the Ullmann-Goldberg reaction. Due to the low complex stability in the Ullmann-Goldberg reaction, only bidentate ligands were analysed due to the additional stability from the chelate effect. The majority of bidentate ligands are N−N, O−O or N−O ligands, with only 7% containing a coordinating sulfur or a phosphate group. Nitrogen-containing functional groups are the most common with amines, amides, imines and pyridines making up the majority of coordinating functional groups, especially in cases where both coordinating functional groups are identical. A summary of the coordinating functional groups present in ***ligands_lit_set*** is shown in **Table 4.1**.

Table 4.1: Frequency of functional groups which coordinate to the copper centre in the ***ligands_lit_set*** for bidentate ligands only. 1 group represents only one of the two coordinating functional groups, 2 groups are both functional groups.

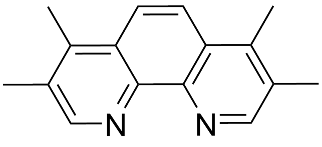
Functional Group	Frequency	1 Group	2 Groups
alcohol	22	6	8
aldehyde	1	1	0
amide (N)	51	33	9
amide (O)	38	38	0
amine	138	48	45
carbene	7	1	3
carboxylic acid	41	39	1
ester	4	4	0
ether	1	1	0
hydrazine	13	11	1
imidazole	5	1	2
imine	41	13	14
indole	4	2	1
ketone	24	8	8
nitrile	2	0	1
N-oxide	15	13	1
oxime	10	6	2
phenol	30	24	3
phosphate	2	2	0
phosphine	12	4	4
phosphine oxide	2	2	0
pyrazole	3	1	1
pyridine	56	30	13
pyrimidine	1	1	0
pyrrole	26	24	1
selenophene	1	1	0
tetrazole	1	1	0
thiol	2	2	0
thiophene	2	2	0
thiophenol	1	1	0
triazole	1	1	0

Importantly, 67% of the bidentate ligand contain a 2-atom bridge, 26% a 3-atom bridge, and 3% a 4-atom bridge. Given the dominance of bidentate ligands with two atoms bridges, they were selected as the preferred mode of coordination for the ligand search.

Table 4.2: Number of bridging atoms between the ligand coordinating atoms for the ligands in the ***ligands_lit_set***.

Number of Bridging Atoms	Frequency
1	2
2	187
3	73
4	9
5+	7

A ligand with a 2 atoms bridge and second-row donor atoms, *i.e.* **TMPHEN**, was selected for generating the template for the transition states which would be employed in the ligand search. **TMPHEN** was also selected due to its limited conformational flexibility, enabling easier identification of transition states, while still being relatively bulky, due to the methyl groups, to provide some steric bulk around the copper centre. **TMPHEN** is also commonly used in the literature so the presence of the transition states is highly likely.



TMPHEN

Figure 4.3: Chemical structure of **TMPHEN**.

4.2.3 Ligand Discovery

Ligands for amine and amide nucleophiles, the two most common coupling partners, were explored. Piperidine (**PIP**) and 2-pyrrolidinone (**PYR**) were selected as coupling partners in order to minimise conformational flexibility in the organometallic intermediates and transition states (**Figure 4.9**). The two non-radical-based mechanisms, oxidative-addition/reductive-elimination (**TSOA**) and sigma metathesis (**TSSig**) were used to calculate the activity of the ligands. Both mechanisms only contain closed-shell transition metal complexes. Radical mechanisms were not included due to the difficulty of modelling open-shell transition metal complexes in a high-throughput manner. For both nucleophiles, the **TSOA** and **TSSig** transition states were generated using **TMPHEN** as the ligand, and iodobenzene as the aryl halide. GFN2-xTB was used to identify the transition states as it's the same level of theory as the high-throughput computational workflow, ensuring the reference structures are as close to the transition state as possible. The structure was optimised to a transition state and the imaginary frequency was checked for the correct vibrational mode. These transition states were used as reference structures to generate catalophores in CSD-CrossMiner to identify potential ligands from the CSD. Reference structures were named ***Pip_TSOA_ref*** and ***Pip_TSSig_ref*** for the **TSOA** and **TSSig** transition states for piperidine and ***Pyr_TSOA_ref*** and ***Pyr_TSSig_ref*** for the **TSOA** and **TSSig** transition states for 2-pyrrolidinone.

As both pathways need to be compared, the same ligand set is required for both pathways. Therefore, the more sterically demanding **TSOA** reference structure was used to generate the catalophores. The less sterically demanding **TSSig** reference structures will yield a higher number of potential ligands, of which a high number will fail due to steric clashes during the generation of the **TSOA** starting structure. A summary of the catalophore generation process for both nucleophiles is shown in **Figure 4.4**.

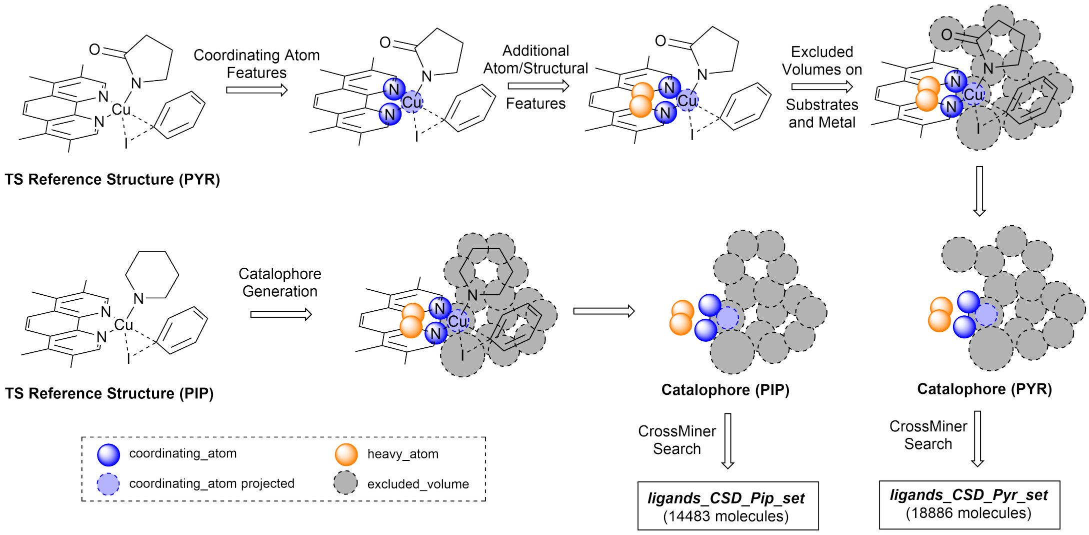
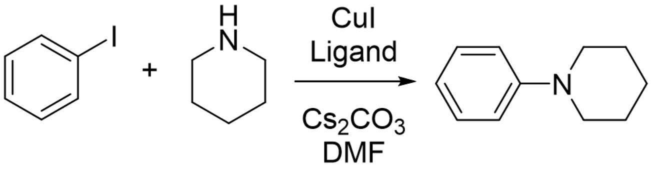

TS Reference Structure (PYR)
Coordinating Atom Features
Additional Atom/Structural Features
Excluded Volumes on Substrates and Metal
TS Reference Structure (PIP)
Catalophore Generation
Catalophore (PIP)
Catalophore (PYR)
CrossMiner Search
CrossMiner Search
ligands_CSD_Pip_set (14483 molecules)
ligands_CSD_Pyr_set (18886 molecules)
coordinating_atom
heavy_atom
coordinating_atom projected
excluded_volume
N
Cu
O
I

Figure 4.4: Workflow for the generation of catalophores for the Ullmann-Goldberg reaction with a piperidine (**PIP**) nucleophile and a 2-pyrrolidinone (**PYR**) nucleophile to yield the datasets ***ligands_CSD_Pip_set*** and ***ligands_CSD_Pyr_set*** respectively.

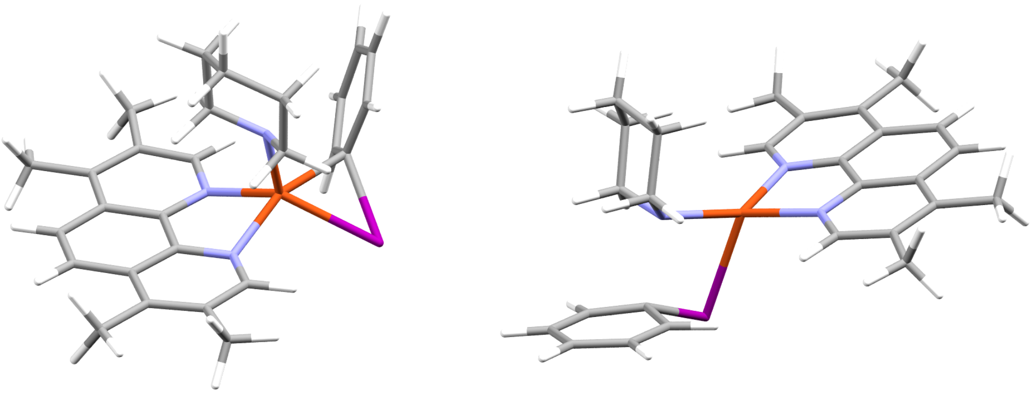
4.2.3.1 Ligand Discovery for Amine Nucleophiles

The ***Pip_TSOA_ref*** reference structure (**Figure 4.5a**) was imported into CSD-CrossMiner and *catsd_coordinating_atom_general* features were placed on each **TMPHEN** nitrogen atom and projected onto the copper atom with a tolerance of 0.75 Å. A bridge of two



Scheme 18: Reaction overview for the Ullmann-Goldberg reaction studied using an amine nucleophile (piperidine).

heavy_atom features between the two nitrogens was placed on the two bridging carbon atoms, with a tolerance of 0.75 Å. A tolerance of 0.75 Å was chosen to define a two-atom bridge while also allowing sufficient flexibility for a three-atom bridge. The features were constrained to be intramolecular. The substrate sites were defined by placing excluded volume features on each atom of piperidine and iodobenzene with a tolerance equal to the van der Waals radii of the base atom (H = 1.20 Å, C = 1.77 Å, N = 1.66 Å and I = 2.04 Å).[204] A smaller tolerance of 1.5 Å was used for copper to allow for coordinating atoms to occupy the space around the copper, while also preventing ligand atoms from occupying the space of the metal. Thus, the created pocket represents the space where both substrates occupy in the transition state of the RDS of the reaction, with soft tolerance allowing the vdW radii of atoms to overlap within the excluded substrate cavity, to allow for variations in individual transition states with different ligands and substrates. Ligands which pre-arrange in this manner will more likely favour the required geometry of the transition state. The catalophore was saved as a *.cm* file (**Figure 4.6**).



(a) TSOA Core

(b) TSSig Core

Figure 4.5: Transition state cores for **TSOA** and **TSSig** for piperidine. Cores were used to generate the catalophore and as a template for structure generation.

The catalophore searches were conducted using the CSD-PythonAPI with a maximum molecular weight of 500 Da, a maximum root-mean-square-deviation (RMSD) in geometry between catalophore and the hit of 1,[205,206] with Br, Cl, I, Li, Na, K, Ca, Mg, Be and transition metals excluded.[93] Only organic structures were included in the search by setting *is_organic* to True. SMILES code matching was used to remove duplicate structures. 3D structures were cleaned by assigning all unknown bond types, adding all missing hydrogens and setting all formal charges.

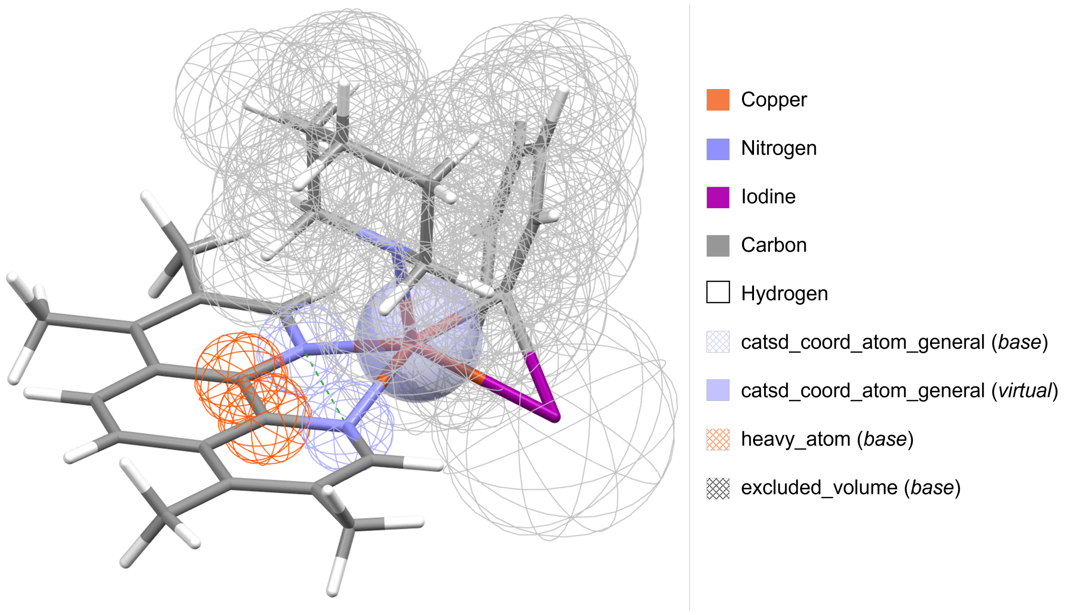
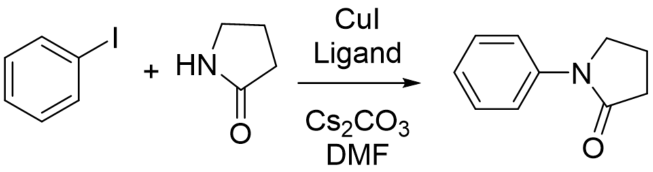


Figure 4.6: Catalphore used for the CSD-CrossMiner search of the CSD the piperidine nucleophile.

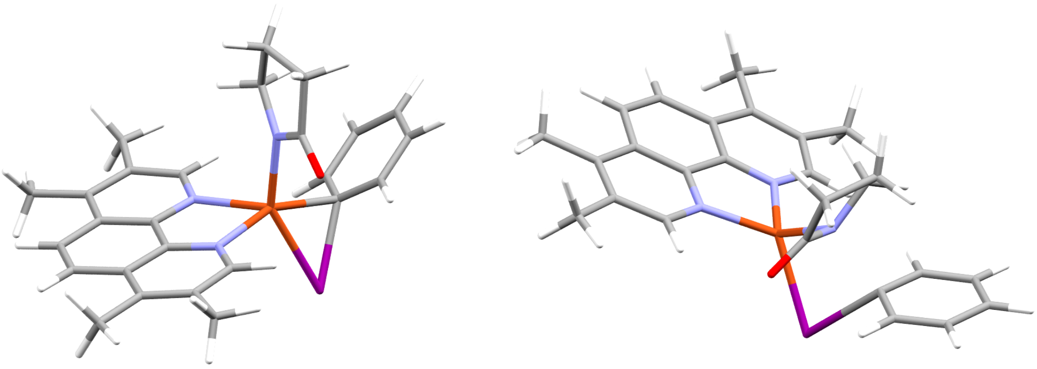
The catalophore search resulted in 26022 total hits, 14483 of which were unique. 27 hits failed due, likely due to issues with the data contained in the CSD and how it is accessed through the Pharmacophore API in the CSD-Python API. Indexes of the coordinating atoms were automatically identified from the hit structure by matching the *catsd_coordinating_-atom_general* feature, used to define the coordinating atoms, to the base atom and exported for use in structure generation. A molSimplify *.dict* file containing all of the relevant data required to use the 3D structures to generate complexes using molSimplify was also generated. Structures were exported in *.mol* format. The resulting ligand set was named ***ligands_CSD_PIP_set***.

4.2.3.2 Ligand Discovery for Amide Nucleophiles



Scheme 19: Reaction overview for the Ullmann-Goldberg reaction studied using an amide nucleophile (2-pyrrolidinone).

The ***Pyr_TSOA_ref*** reference structure (**Figure 4.7a**) was imported into CSD-CrossMiner and *catsd_coordinating_atom_general* features were placed on each **TMPHEN** nitrogen atom and projected onto the copper atom with a tolerance of 0.75 Å. A bridge of two *heavy_atom* features between the two nitrogens was placed on the two bridging carbon atoms, with a tolerance of 0.75 Å. The features were constrained to be intramolecular. The substrate sites were defined by placing excluded volume features on each atom of 2-pyrrolidinone and iodobenzene with a tolerance equal to the van der Waals radii of the base atom (H = 1.20 Å, C = 1.77 Å, N = 1.66 Å, O = 1.50 Å and I = 2.04 Å). A smaller tolerance of 1.5 Å was used for copper to allow for coordinating atoms to occupy the space around the copper, while also preventing ligand atoms from occupying the space of the metal. The catalophore was saved as a *.cm* file (**Figure 4.8**).



(a) TSOA Core **(b)** TSSig Core

Figure 4.7: Transition state cores for **TSOA** and **TSSig** for 2-pyrrolidonone. Cores were used to generate the catalophore and as a template for structure generation.

The catalophore searches were conducted using the CSD-PythonAPI with a maximum molecular weight of 500 Da, a maximum root-mean-square-deviation (RMSD) in geometry between catalophore and the hit of 1,[205,206] with Br, Cl, I, Li, Na, K, Ca, Mg, Be and transition metals excluded.[93] Only organic structures were included in the search by setting *is_organic* to True. SMILES code matching was used to remove duplicate structures. 3D structures were cleaned by assigning all unknown bond types, adding all missing hydrogens and setting all formal charges.

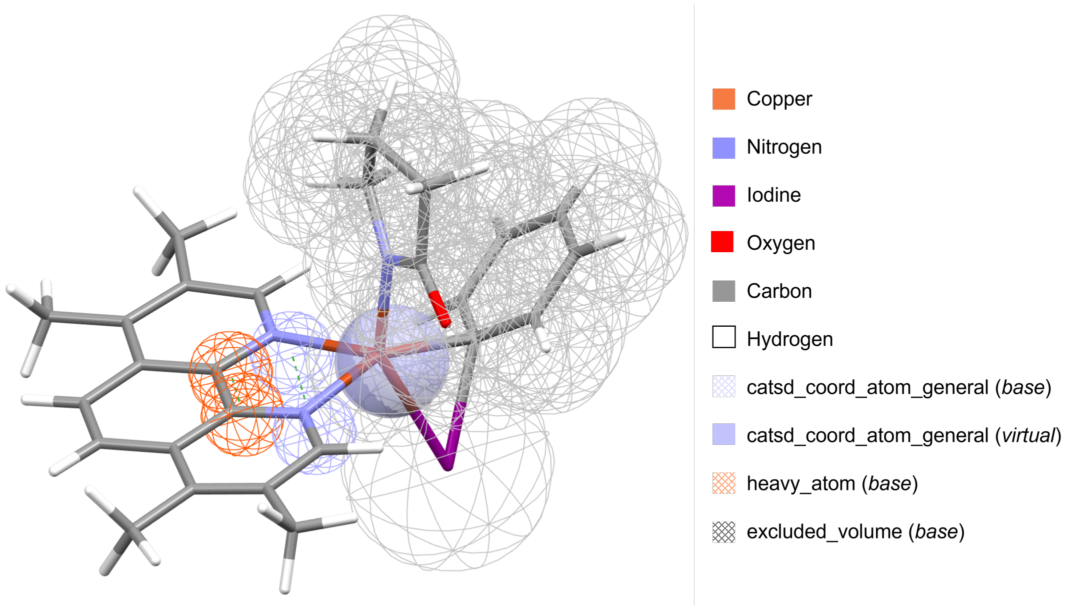

🟧 Copper
🟦 Nitrogen
🟪 Iodine
🟥 Oxygen
⬛ Carbon
□ Hydrogen
▨ catsd_coord_atom_general (*base*)
▨ catsd_coord_atom_general (*virtual*)
▨ heavy_atom (*base*)
▨ excluded_volume (*base*)

Figure 4.8: Catalphore used for the CSD-CrossMiner search of the CSD for the 2-pyrrolidinone nucleophile.

The catalophore search resulted in 33780 total hits, 18886 of which were unique. 37 hits failed due, likely due to issues with the data contained in the CSD and how it is accessed through the Pharmacophore API in the CSD-Python API. Indexes of the coordinating atoms were automatically identified from the hit structure by matching the *catsd_coordinating_-atom_general* feature, used to define the coordinating atoms, to the base atom and exported for use in structure generation. A molSimplify *.dict* file containing all of the relevant data required to use the 3D structures to generate complexes using molSimplify was also generated. Structures were exported in *.mol* format. The resulting ligand set was named ***ligands_CSD_PYR_set***.

4.2.4 Calculation of Activation Energies

Activation energies were calculated for a set of commonly used synthetic conditions. Copper(I) iodide was used as the precatalyst, caesium carbonate as the base and N,N-dimethylformamide (DMF) as the solvent. Two non-radical pathways were explored, oxidative addition/reductive elimination, **TSOA**, and sigma metathesis, **TSSig**, (**Figure 4.9**).

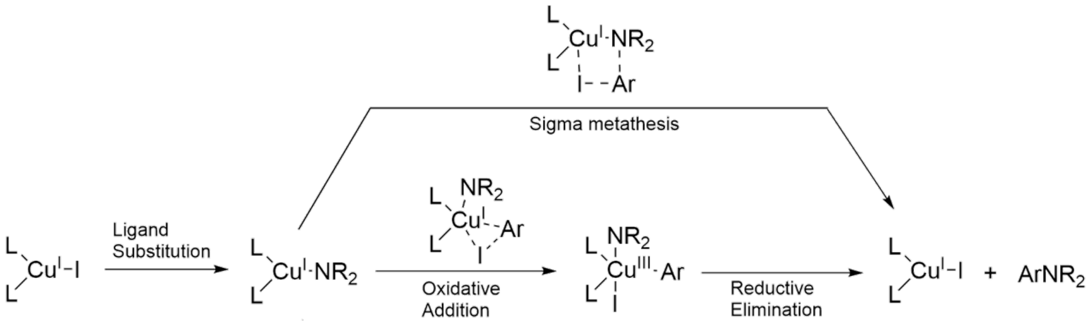


Figure 4.9: The mechanisms of the Ullmann-Goldberg reaction explored, containing important intermediates and transition state structures.

Four complexes were generated for each pathway. The precatalyst after ligand exchange and product complex, **CuLI**. The active catalytic species, containing one bidentate ligand and the deprotonated nucleophile coupling partner, **CuLNu**. For the oxidative addition pathway, the oxidative addition step is the rate-determining step, therefore only this transition state is generated. For the sigma metathesis pathway, only one transition state is required. The stable intermediates, **CuLI** and **CuLNu** are common across both pathways, therefore, only four complexes need to be generated for each ligand.

4.2.4.1 Ligand protonation state rules

As organic N, O and S based ligands often contain alpha hydrogens, the protonation state of the ligand coordinating atoms needs to be considered. Protonation or deprotonation of the ligand results in a different charge on the complex, electronic properties and differences in geometries. The presence of a hydrogen atom will significantly alter the steric environment around the copper due to the difference in steric bulk between sp^2/trigonal planar and sp^3/tetrahedral nitrogen atoms. Deprotonation also influences bonding properties as a free nitrogen lone pair in the p orbital of N$^-$ can provide significant pi donation to the metal centre due to better orbital overlap with the copper d orbitals compared to neutral tetrahedral nitrogen, where the lone pair is used in the Cu$-$N σ-bond.

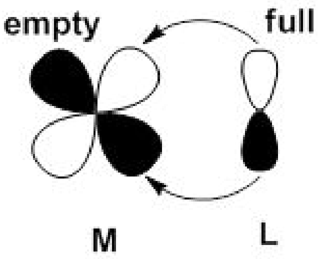


Figure 4.10: Pi-donation of a full p-orbital with metal d-orbitals.

A set of ligand protonation state rules was generated via analysis of the protonation states of common ligand coordinating functional groups in the Cambridge Structural Database. The number of protonated and deprotonated ligands available for each functional group was extracted for complexes with only one copper atom. The results are shown in **Table 4.3**.

Table 4.3: Protonation states of common ligand functional group from the Cambridge Structural Database and their associated pKa range in DMSO.[207]

Functional Group	Number of entries	Unchanged	Deprotonated	pKa[207]
Amine	3793	3747	46	~40
Aniline	199	165	34	25–31
Hydrazine	102	101	1	25–29
Imine	218	215	3	~31
Amide	906	14	882	17–25
Carboxylic acid	4622	18	4604	9–13
Alcohol	1169	991	178	~30
Thiol	34	1	33	5–12
Phenol	2939	99	2840	10–19
Thiophenol	68	0	68	5–12
Phosphonic acid	209	4	205	~2

Nitrogen-based donors tend to keep their hydrogen atom, with amines, anilines, hydrazines and imines all having a high ratio of protonated to deprotonated structures. The only exception is amides where the deprotonated donor is favoured due to resonance stabilisation of the negative charge with the oxygen atom. There are several exceptions where there are other functional groups present in the ligand that are able to stabilise the generated negative charge. Deprotonated ligands are also observed for carboxylic acids, phenols and thiophenols, all of which possess a high degree of resonance stabilisation. Alcohols favour the protonated form upon complexion with copper, with a 9:1 ratio of protonated:deprotonated structures, whereas thiols favour the deprotonated form, this is likely due to the 'soft' diffuse nature of the sulfur atom. Phosphorous-containing O-donors also show a high preference for deprotonation due to resonance. Analysis of functional group pKa's suggests that those functional groups with a pKa $<\sim 25$ in DMSO are deprotonated upon complexation whereas those with a pKa $>\sim 25$ remain protonated (see reference for a collection of pKa values).[207] A set of deprotonation rules for forced hydrogen removal was generated based on functional group SMILES strings (**Table 4.4**).

Table 4.4: SMILES strings and associated functional group for the protonation rules used for structure generation.

Functional Group	SMILES String
Amide	O=CN
Carboxylic Acid	O=CO
Aromatic nitrogen	n
Amidine	N=CN
N-amino	nN
Phenol	Oc
Thiophenol	Sc
Thiol	SC
N-oxide	ON
Aromatic N-oxide	On
Sulphonic acid	OS
N-Carbene	NCN
Phosphonic	OP

SMILES strings for commonly deprotonated groups such as N-oxides, sulphonic acids and carbenes are included to remove any uncertainty due to implicit hydrogen atoms in the SMILES string or for 3D structures which contain a protonated functional group. This set of rules dictates automatic deprotonation of the ligand during complex generation using molSimplify.

4.2.4.2 Complex Generation

Organometallic complexes were generated with the molSimplify Python toolkit as described in **Section 3.2.5**. Intermediate structures, **CuLI** and **CuLNu** were generated using a copper oxidation state of (I), trigonal planar geometry and spin 1 ($4s^0 3d^{10}$). Nucleophile structures were optimised separately using GFN2-xTB and supplied as an *.xyz* file. Ligands were supplied using the *.mol* 3D structure retrieved from the CSD. The smart alignment method (*ligalign*) was enabled to minimise steric interactions and the complex was optimised using the universal force field before and after addition. Forced hydrogen removal was used to deprotonate the amine/amide nucleophile, and automatic removal was applied to the ligands using SMILES string matching with the deprotonation rules presented in **Section 4.2.4.1** by modification of the source code (described in **Section 3.2.5.1**). Only one bidentate ligand was added to ensure a three-coordinate, trigonal planar geometry. These structures were used as starting guesses for quantum mechanical geometry optimisation.

In order to automate the generation of transition states **TSOA** and **TSSig**, a different strategy was employed. Reference structures, ***Pip_TSOA_ref***, ***Pip_TSSig_ref***, ***Pyr_TSOA_ref*** and ***Pyr_TSSig_ref*** generated with **TMPHEN** were used as the core. The **TMPHEN** ligand was substituted with the ligand of interest via ligand replacement (described in **Section 3.2.5.2**) using the same 3D structure from the CSD. An oxidation state of 0 was used (the charge on the core) with spin 1. The smart alignment method (*ligalign*) was enabled to minimise steric interactions and automatic hydrogen removal was applied to the ligands using the same deprotonation rules. Complexes were then optimised with the custom After-Core Constrained method using the Universal Force Field (UFF), where the transition state ‘core’ is locked and only the ligand is optimised to ensure the transition state mode is preserved.

4.2.4.3 Tuning the Vetting Procedure

The automatic structure checking criteria, described in **Section 3.2.4.1**, were tuned to improve the success rate of the automatic identification procedure. Intermediate structures, **CuLI** and **CuLpip/CuLpyr** were confirmed to be at a minimum (zero imaginary frequencies). For transition states the structure was confirmed to have one imaginary frequency. A cutoff value of -40 cm^{-1} was used for the imaginary frequency, as any imaginary frequency between -40 and 0 could be considered to be numerical noise. The C−I and Cu−C bonds for **TSOA**, and the C−N bond for **TSSig**, were checked to be at an intermediary length. Covalent radii for Cu, N, C and I were used to determine the default bond length.[208]

As the C−I and Cu−C bonds in the oxidative addition transition state are not a proper bond stretch and possess a bending component the value of S_0 had to be tuned. The value of S_0 was decreased in increments of 0.05 until there was a decrease in accuracy. 198 transition states from the ***ligands_CSD_PIP_set_TSOA*** were used as a test set.

Table 4.5: Successful identification of a correct transition state with varying values of S_0 for 198 ligands in the TSOA_PIP dataset. The 'Not used' entry only uses the magnitude of the imaginary frequency to determine if the transition state is correct.

S_0	Correctly Identified Transition States	Incorrectly Identified Transition States	Success Rate (%)
Not used	175	23	88.4
0.20	197	1	99.4
0.25	180	18	90.9
0.30	142	56	71.7
0.33	114	84	57.6

The success rates for a range of values of S_0 are shown in **Table 4.5**. The default value, 0.33, is not sufficient for the oxidative addition transition state, successfully identifying only 57.6% of the transition states correctly. Successful identification of correct transition states improved by decreasing the value of S_0 until a value of 0.20 (99.4%). Decreasing the value of S_0 below 0.20 resulted in a large increase in false positives. Therefore, 0.20 was used as the value of S_0 for the vetting of all oxidative addition transition states. As the C−N bond in the sigma metathesis transition state can be considered a proper stretch along the C−N bond the default value of 0.33 was used for S_0.

4.2.4.4 Computational Calculations

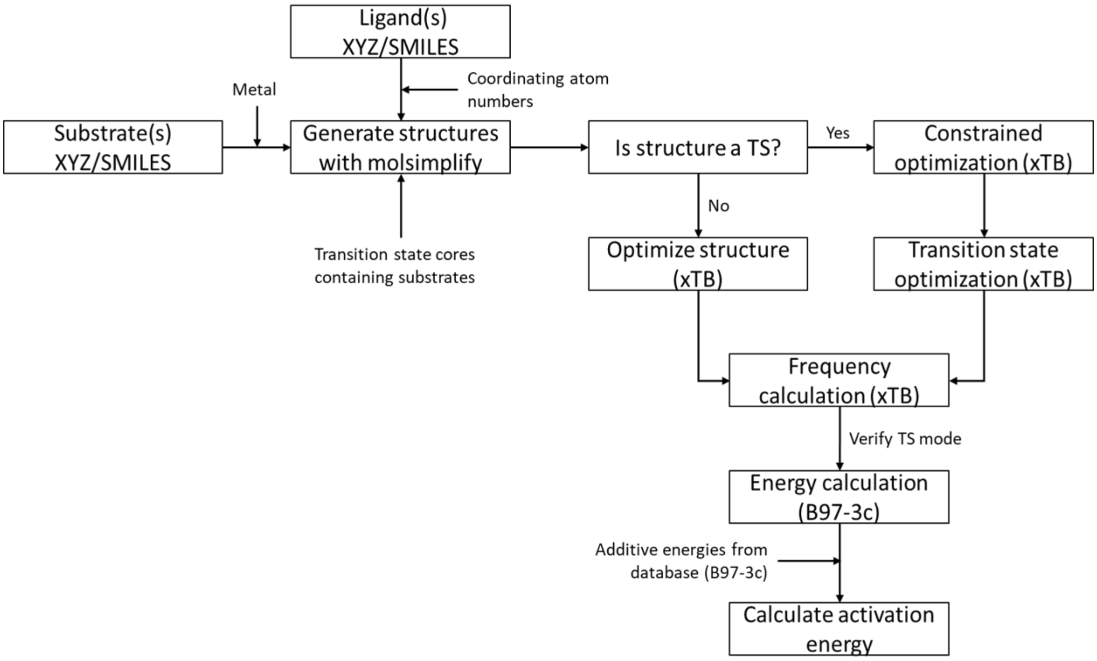

Ligand(s)
XYZ/SMILES
Coordinating atom numbers
Metal
Substrate(s)
XYZ/SMILES
Generate structures with molsimplify
Transition state cores containing substrates
Is structure a TS?
Yes
No
Constrained optimization (xTB)
Optimize structure (xTB)
Transition state optimization (xTB)
Frequency calculation (xTB)
Verify TS mode
Energy calculation (B97-3c)
Additive energies from database (B97-3c)
Calculate activation energy

Figure 4.11: Overview of the workflow presented in **Chapter 3** for the high-throughput calculation of activation energies.

All structures were optimised using the procedure described in **Chapter 3**, which is summarised in **Figure 4.11**. The complexes were generated using molSimplify and pre-optimised with GFN2-xTB using the *TightOpt* optimisation criteria. For transition states the bond lengths and bond angles of the atoms involved in the imaginary frequency were frozen. The C−I, Cu−I and Cu−C bonds and I−Cu−C bond angle were frozen for **TSOA**. The C−N and C−I bond lengths and N−C−I bond angle were frozen for **TSSig**. Transition states were optimised with GFN2-xTB with recalculation of the hessian every five optimisation steps. The presence of the correct transition state is verified using the automated structure validation criteria (**Section 3.2.4.1**). Single point energies were calculated at the B97-3c level of theory using the *TightSCF* SCF convergence criteria and the *SlowConv* convergence method. No conformational search was undertaken as component structures are all taken from 3D X-ray structures (ligands) or optimised minima (nucleophiles). This allows for a consistent comparison between ligand conformations. DMF was used as the solvent for all calculations.

The piperidine pre-optimisation calculations with GFN2-xTB were performed in the standalone xtb 6.3.3 program. For the 2-pyrrolidinone, all GFN2-xTB pre-optimisation calculations are performed in ORCA interfaced with xtb 6.3.3 to ensure consistency with the optimiser used. ORCA interfaced with xtb 6.3.3 is used for all transition state optimisations for both nucleophiles. Using the same optimiser for both the pre-optimisation and transition state optimisation ensures that in the starting structure for the transition state optimisation, the ligand is as close to a minimum as possible within the optimiser. Using an external optimiser can introduce slight negative frequencies reducing the reliability and increasing the computational time required to reach the saddle point.

Additives (e.g. base, starting materials and products) were calculated separately. Structures were optimised at the GFN2-xTB level of theory and single-point energies were calculated at the B97-3c level of theory. DMF was used as the solvent in all cases. Gibbs free energies were calculated and stored in a database.

4.2.5 Data Analysis

4.2.5.1 Success Rates

Initial structure generation resulted in successful complex generation for 14451 ligands with only 32 ligands failing for the ***ligands_CSD_PIP_set*** dataset. For ***ligands_CSD_PYR_-set***, correct complexes were generated for 18848 ligands with only 38 ligands failing. The generation of initial structures had a success rate of 99.7% for both datasets. This demonstrates the applicability of the catalophore search term for finding suitable candidate ligand structures which fit around the metal centre for a specific set of substrates.

When the computational workflow was applied to ***ligands_CSD_PIP_set*** and ***ligands_-CSD_PYR_set***, a significant decrease in optimisation success rates were observed for both stable intermediates and transition states. While these are expected due to the level of complexity in these ligand candidates, the success rate of finding and optimising **TSOA** (33%) was particularly low using ***ligands_CSD_PIP_set*** (**Table 4.6**). However, this result may simply reflect that many potential ligands are not suitable for the Ullmann-Golberg reaction, as suggested by the experimental literature, or that oxidative addition is not the correct mechanism for the majority of ligands or this set of substrates.

Table 4.6: Structure generation and optimisation (inside bracket) success rates for the high-throughput calculations for both nucleophiles.

Structure	***ligands_lit_set*** Bidentate with **Pip** only	***ligands_CSD_Pip_set***	***ligands_CSD_Pyr_set***
CuLI	98(93)	99(85)	99(84)
CuLpip/CuLpyr	98(93)	99(77)	99(89)
TSOA	97(88)	99(33)	99(61)
TSSig	99(94)	99(85)	99(83)

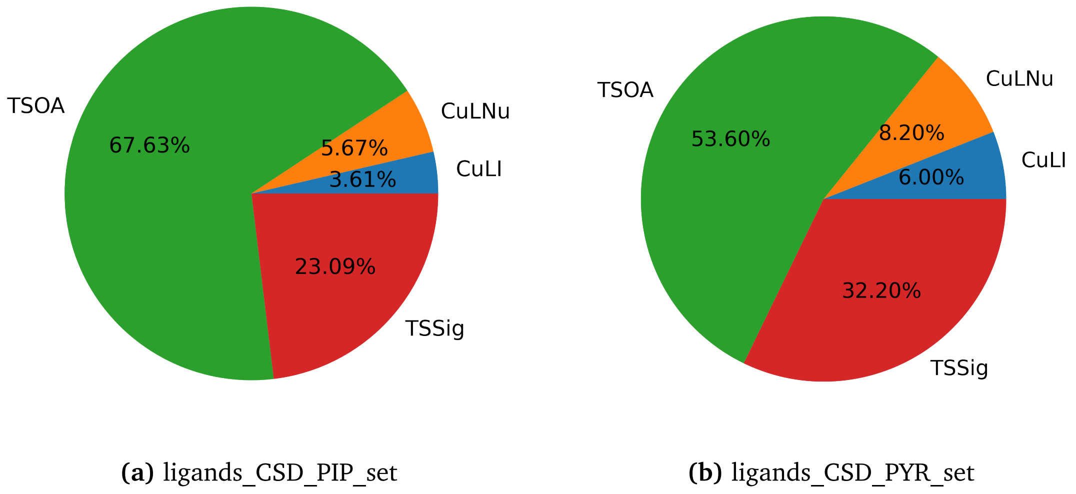
The most common cause for intermediate structure failing is the structures having small imaginary frequencies (-40 $< x <$ 0 cm^{-1}). This could be improved by using tighter convergence criteria, however, as the *Tight* convergence criteria was already in use, increasing the criteria to *VeryTight* is a significant increase in computational time to improve only 7% of structures. Therefore, these structures were not rerun. Another approach to improving the success rate is using an iterative optimisation method, whereby if a negative frequency is identified, the structure is distorted along the imaginary mode and reoptimised. However, the iterative optimisation method in ORCA is not compatible with GFNx-xTB methods at the time of development. When this becomes available it should provide increased reliability for non-transition state structures.

The improvement in success rate for the **CuLNu** intermediate from 77% to 89% is likely due to the change to using the ORCA optimiser for the optimisation. As frequency calculations are all calculated in ORCA, it is likely that the difference in optimiser resulted in small imaginary frequencies in ORCA. 12.8% of all ligands fail for both transition states for the ***ligands_CSD_PIP_set***. A similar value is observed for ***ligands_CSD_PYR_set*** with 11.3% of ligands failing for both transition states. This suggests that approx. 12% of ligands in both datasets are unsuitable as ligands in the Ullmann-Goldberg reaction. The oxidative addition transition state has a much lower success rate than the sigma metathesis transition state for piperidine with 9663 and 2215 ligands failing respectively for ***ligands_CSD_-PIP_set***. The increase in **TSOA** success rate from 33% to 61% for ***ligands_CSD_PYR_set*** suggests that the oxidative addition transition state is more viable with an amide nucleophile compared to an amine. A small percentage of improvement may be attributed to the change in pre-optimisation from xtb to ORCA however, the introduction of several small imaginary frequencies is unlikely to cause such a large increase in success rate. The sigma metathesis transition state has a much higher success rate than the oxidative addition transition state across both datasets. This suggests that the sigma metathesis transition state is a more viable transition state for the Ullmann-Goldberg reaction.

4.2.5.2 Computational Time

Structures were generated on a 2-core laptop in series taking 3.9 days for the generation of 57,696 structures for ***ligands_CSD_PIP_set*** and 5.7 days for the generation of 75244 structures for ***ligands_CSD_PYR_set***. The average time taken to generate one complex is 6 seconds. The use of parallelisation should dramatically speed up structure generation if available.

The computational time for each calculation was extracted from the output files and converted into a single-core time. The real-world runtime on a high-performance computer using 4 cores and 1GB of RAM per calculation for ***ligands_CSD_PIP_set*** is ∼6 weeks for 14483 ligands. The comparative time for ***ligands_CSD_PYR_set*** is ∼4 weeks for 18886 ligands. The breakdown of the time taken per complex for each ligand is shown in **Figure 4.12**.



(a) ligands_CSD_PIP_set **(b)** ligands_CSD_PYR_set

Figure 4.12: Breakdown of computational time for each complex in the mechanistic pathway as a percentage of the total time for all calculations.

Transition state calculations compose the majority of the computational resources required per ligand, constituting 90.2% of the total computational time for ***ligands_CSD_PIP_set*** and 85.8% of the total computational time for ***ligands_CSD_PYR_set***. The proportion of time taken for the **TSOA** calculation decreases for the ***ligands_CSD_PYR_set*** dataset. This is likely due to the increased success rate of the **TSOA** calculations. Optimisations that do not converge to the correct transition state often take much longer due to the increased number of steps required to move away from the starting structure to an incorrect transition state, especially with frequent recalculation of the hessian (∼20 minutes per recalculation). A full breakdown of the single-core computational time for each calculation, as well as the average time in hours for each ligand, is shown in **Table 4.7**.

Table 4.7: Breakdown of single-core computational time in hours for each calculation for all ligands in the ***ligands_CSD_PIP_set*** and ***ligands_CSD_PYR_set*** datasets. All values except for the average are reported as the sum for all ligands.

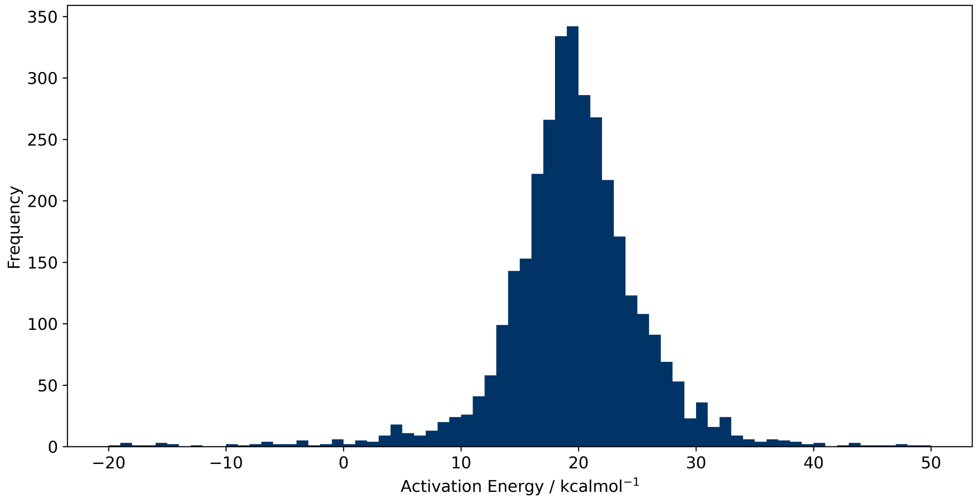
Structure	Single-Core Computational Time (h)					
	Preoptimisation	Optimisation	Frequency	Energy	Total	Average (per ligand)
ligands_CSD_PIP_set						
CuLI	-	66	3176	3068	6310	0.44
CuLpip	-	169	5058	4839	10 067	0.70
TSOA	224	102 747	7175	8189	118 336	8.20
TSSig	210	30 619	7155	7802	45 787	3.17
ligands_CSD_PYR_set						
CuLI	-	1173	4359	4237	9789	0.52
CuLpyr	-	1693	6109	5509	13 312	0.71
TSOA	1946	67 266	8737	9212	87 162	4.63
TSSig	2513	30 740	8773	10 439	52 466	2.79

The average time for all calculations per ligand is 12.6h for ***ligands_CSD_PIP_set*** and 8.7h for ***ligands_CSD_PYR_set***. This is a similar time scale to an analogous experimental reaction, not accounting for the synthesis of the starting materials or ligands. All intermediate structures take less than 1h per ligand with the **CuLI** complex taking approximately 30 minutes per ligand, with the majority of computational time taken by the frequency and energy calculations. Frequency and energy calculations scale well to the transition states taking a similar time scale when accounting for the added number of atoms in the calculation. The limiting factor in regards to computational time for the transition states is the transition state optimisation calculation taking 2.9-4.3× longer for **TSSig** and 7.7-14.3× longer for **TSOA** than the frequency calculation. Across both datasets, the **TSOA** calculations take longer than **TSSig**. This is likely due to the potential energy surface being shallow, taking more optimisation steps to reach the transition state. It could also be due to the increased failure rate of these calculations.

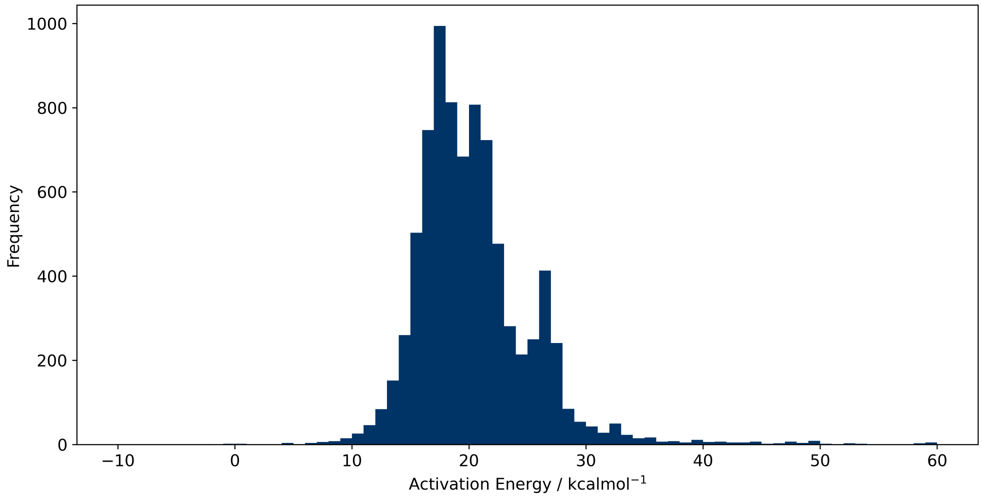
Moving from preoptimisation in xtb to ORCA shows a small improvement in the average computational time for the **TSSig** transition state from 3.17h to 2.79h per ligand. The total computational time stays almost the same for the transition state optimisation step but with ∼4000 more ligands calculated. A significant improvement is seen for **TSOA** taking 0.65× the computational time for the transition state optimisation step for 1.3× more ligands. Per ligand, the time taken reduces from 8.2h to 4.6h saving 3.6h per ligand. While the ORCA optimiser takes almost 6.5× as long for the preoptimisation step (accounting for the increased number of calculations), the total time saved from the transition state optimisation is large (66,300h for **TSOA**, 9,000h for **TSSig**). This demonstrates that the ORCA optimiser is superior for providing a better starting structure for the transition state optimisation, reducing the total number of steps needed to reach the transition state.

4.2.5.3 Activation Energies

The activation energies for all ligands in the ***ligands_CSD_PIP_set*** and ***ligands_CSD_PYR_set*** datasets were calculated automatically for both pathways using Python. Additive energies were taken from the database of additive energies calculated previously. Each dataset was filtered to remove ligands with incorrect structures (e.g. incorrect transition state or non-minimum intermediate) and separated into separate datasets for **TSOA** and **TSSig**. This resulted in four datasets, ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig***. The activation energy distributions for each dataset are shown in **Figures 4.13** and **4.14**. Distributions were truncated at low and high activation energies for clarity. Full plots are available in **Appendix 4.A.3**.



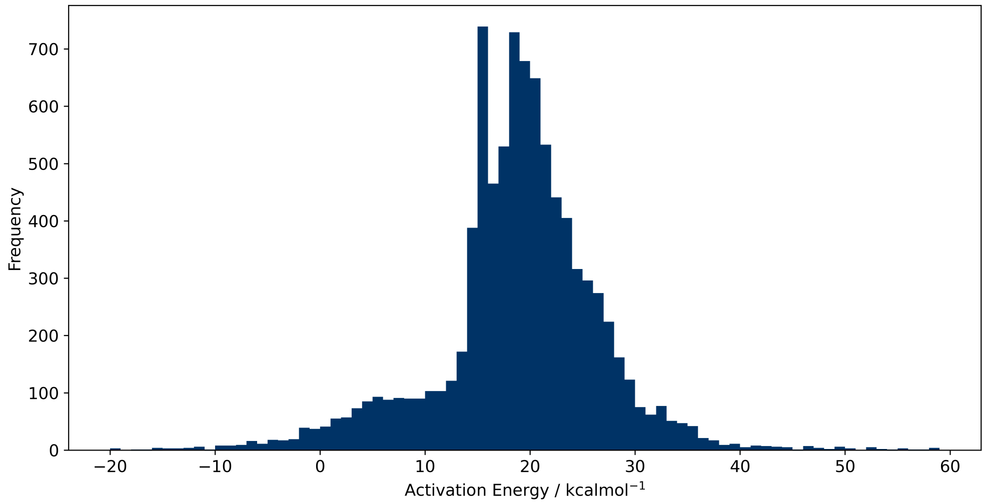
(a) TSOA



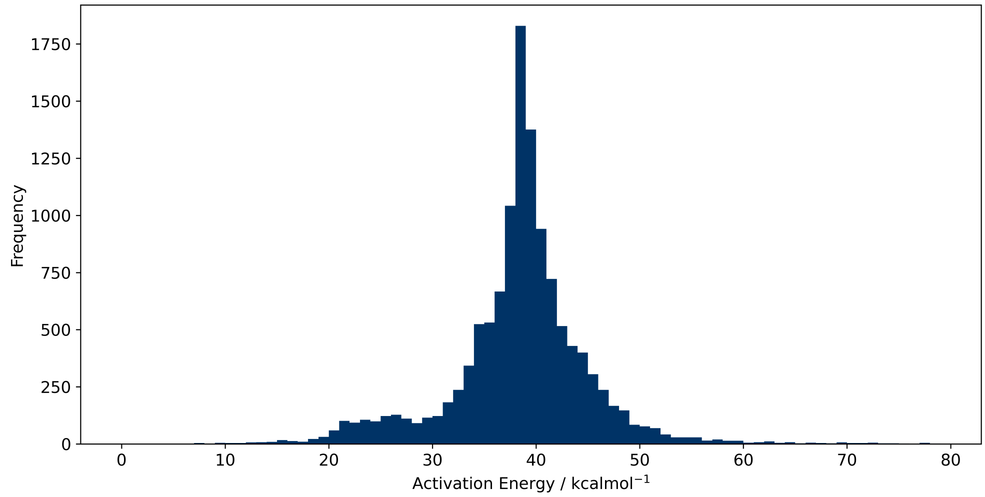
(b) TSSig

Figure 4.13: Distribution of activation energies for both the **TSOA** and **TSSig** transition state for piperidine. Only -20 to 50 kcal mol^{-1} is shown for **TSOA** and -10 to 60 kcal mol^{-1} for **TSSig**.

The activation energy distributions for ***ligands_CSD_PIP_set*** are similar across both transition states with an average $\Delta G^{\ddagger}$ of ~18 kcal mol^{-1}, with the majority of ligands lying between 10 and 30 kcal mol^{-1}. The **TSOA** transition state has a significant number of ligands with a $\Delta G^{\ddagger} < 0$ kcal mol^{-1}. This suggests that the oxidative addition and sigma metathesis pathways are very close in energy for an amine nucleophile.



(a) TSOA



(b) TSSig

Figure 4.14: Distribution of activation energies for both the **TSOA** and **TSSig** transition state for 2-pyrrolidinone. Only -20 to 60 kcal mol^{-1} is shown for **TSOA** and 0 to 80 kcal mol^{-1} for **TSSig**.

Unlike piperidine, the activation energy distributions for ***ligands_CSD_PYR_set*** are different between transition states. The ***ligands_CSD_PYR_set_TSOA*** dataset has an average $\Delta G^{\ddagger}$ of ∼18 kcal mol^{-1}, whereas the ***ligands_CSD_PYR_set_TSSig*** datasets has a higher average $\Delta G^{\ddagger}$ of ∼38 kcal mol^{-1}. This suggests for an amide nucleophile the oxidative addition pathway is energetically more favourable than the sigma metathesis pathway. As with piperidine the **TSOA** transition state has a significant number of ligands with a $\Delta G^{\ddagger}$ < 0 kcal mol^{-1}. Manual inspection of several **TSOA** transition states with negative $\Delta G^{\ddagger}$ across both nucleophiles identified that in most cases a hydrogen atom was transferred between the ligand and the nucleophile nitrogen atom or there is hydrogen bonding present between ligand and the nucleophile nitrogen atom (**Figure 4.17**). All negative activation energy **TSOA** transition states were analysed to determine how many contained a protonated nucleophile nitrogen atom. The results are shown in **Figures 4.15** and **4.16** for piperidine and 2-pyrrolidinone respectively.

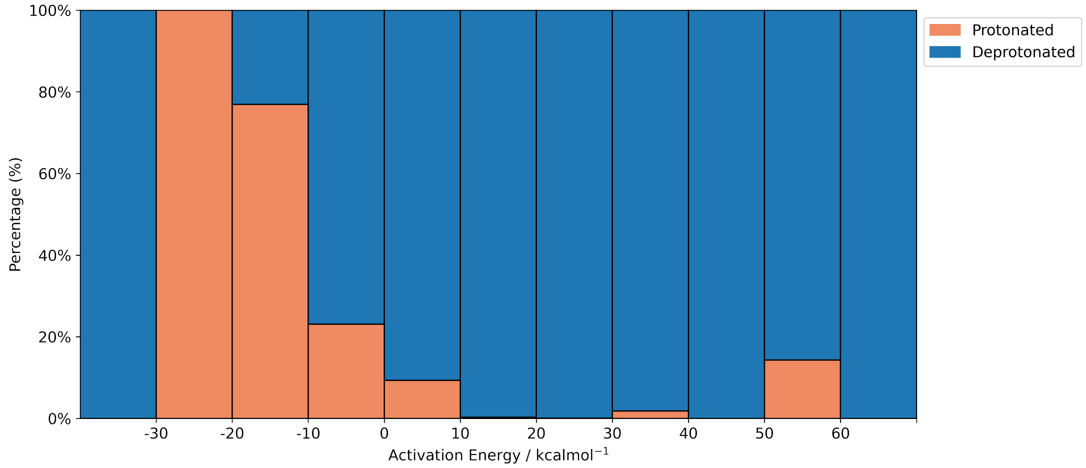


Figure 4.15: Percentage of **TSOA** structures with a protonated piperidine nitrogen in ***ligands_CSD_PIP_set_TSOA***. Transition states are binned by activation energy.

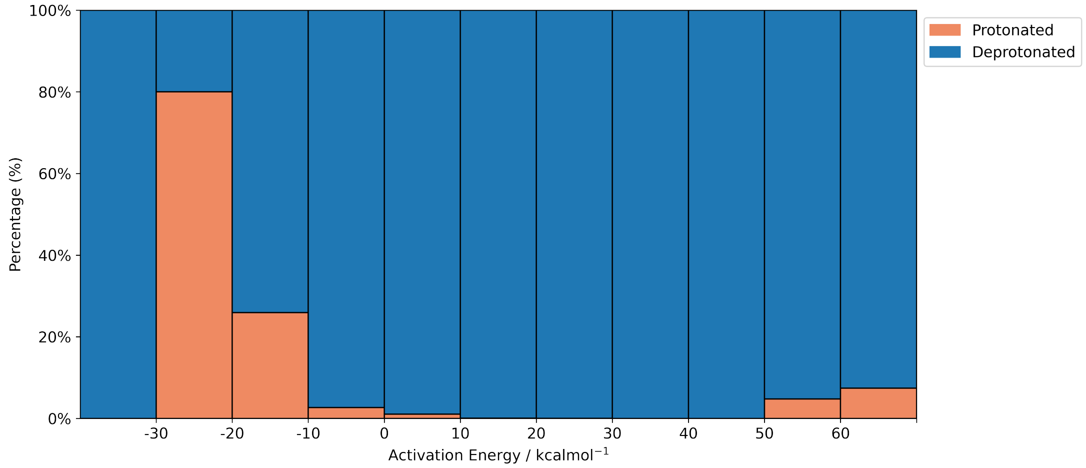
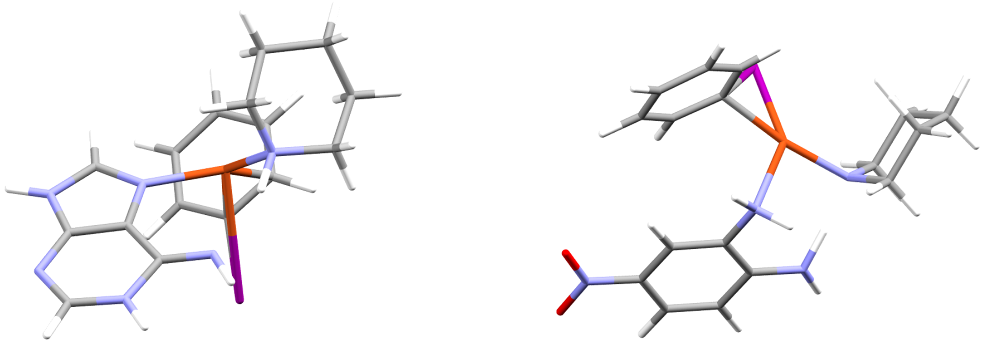


Figure 4.16: Percentage of **TSOA** structures with a protonated 2-pyrrolidinone nitrogen in ***ligands_CSD_PYR_set_TSOA***. Transition states are binned by activation energy.

The majority of structures below 0 kcal mol^{-1} contain a protonated nucleophile for both datasets. Lower activation energies contain a higher percentage of protonated structures. This suggests that protonation of the nucleophile stabilises the oxidative addition step. However, protonation of the nucleophile inhibits the subsequent reductive elimination step, meaning these structures could be deemed incorrect as the protonation state of the ligand is incorrect. The incorrect protonation state results in a transfer of the hydrogen atom from the ligand-coordinating atom to the nucleophile nitrogen, whereas in basic solution the hydrogen would have been deprotonated by the base. There are multiple ligands in both datasets that have a negative activation energy and contain a deprotonated nucleophile.



(a) Hydrogen Atom Transfer **(b)** Hydrogen Bonding

Figure 4.17: Structural features resulting in negative calculated activation energies.

4.2.5.4 Validation of Negative Activation Energies

The ligands with negative activation energies were validated by recalculating the activation energy with a higher-level method. Activation energies were recalculated at the DLPNO-CCSD(T)/def2-TZVPP level of theory and compared with the B97-3c activation energy for the ***ligands_CSD_PIP_set*** dataset. The results are shown in **Figures 4.18** and **4.19**.

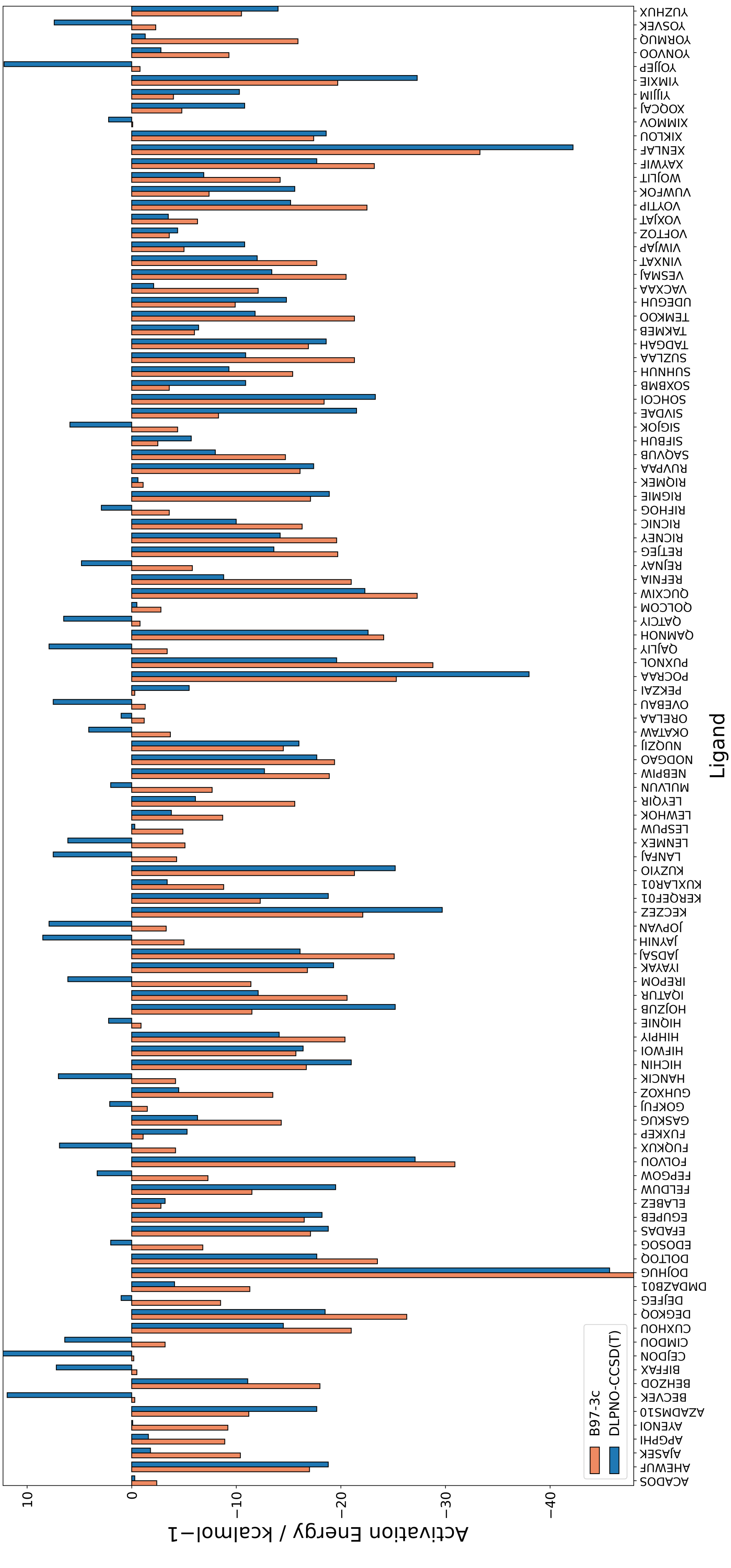


Figure 4.18: Comparison of activation energies calculated with B97-3c (orange) and DLPNO-CCSD(T)/def2-TZVPP (blue) for the **TSOA** transition state with negative B97-3c activation energies.

The majority of activation energies remain negative with only 28 activation energies having positive activation energies at the DLPNO-CCSD(T)/def2-TZVPP level of theory. Of these 28 structures, the majority contain 5-membered O−Cu−O coordinating rings or a 4-membered coordination geometry. This suggests that B97-3c is poor at calculating accurate energies for these motifs. This is consistent with the previous observation during benchmarking where the same structural motifs resulted in poor energies compared to DLPNO-CCSD(T)/def2-TZVPP (see **Section 3.2.7.1**). Of the remaining structures, all activation energies remain negative. This verifies that protonation of the nucleophile or hydrogen bonding between the ligand and the nucleophile can stabilise the oxidative addition transition state. Only 27 ligands with negative activation energies were deemed correct upon manual inspection.

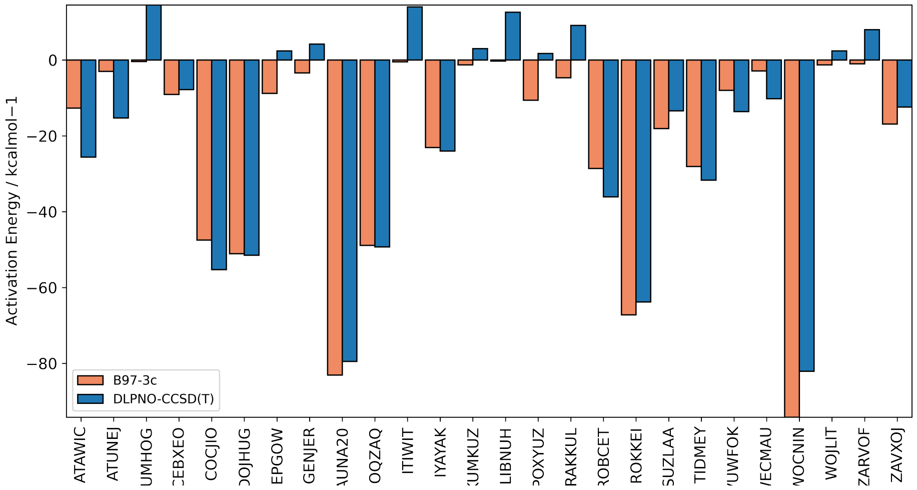


Figure 4.19: Comparison of activation energies calculated with B97-3c (orange) and DLPNO-CCSD(T)/def2-TZVPP (blue) for the **TSSig** transition state with negative B97-3c activation energies. **GUTZAW** was excluded for clarity. (B97-3c = -267.3 kcal mol^{-1}, DLPNO-CCSD(T) = -251.3 kcal mol^{-1}).

For **TSSig** three structures, **BANVUK**, **COJMUJ** and **SINZEW**, had incorrect energies at the B97-3c level of theory. Recalculation at the DLPNO-CCSD(T)/def2-TZVPP level of theory resulted in reasonable positive activation energies of 32.8 kcal mol^{-1}, 9.9 kcal mol^{-1} and 19.3 kcal mol^{-1} respectively. Five ligands had calculated activation energies <-1000 kcal mol^{-1} caused by errors in calculations which were also observed in the DLPNO-CCSD(T) calculations. All five of these structures were deemed incorrect and therefore discarded. Ten ligands had a positive activation energy at the DLPNO-CCSD(T)/def2-

TZVPP level of theory. As with the oxidative addition transition state structures contained 5-membered O−Cu−O coordinating rings or 4-membered coordination geometry. 16 ligands had negative activation energies. These structures contain unusual coordination geometry, hydrogen bonding, protonation of the nucleophile or incorrect structures which were not identified during structure checking.

4.2.6 Experimental Validation

The experimental work presented in this section was performed by our Master's student Zeshen Wang (ligand synthesis) and Dr Tom Nicholls (ligand testing) from the Willans group at the University of Leeds.

To validate the predictions of the computational workflow, 15 ligands were chosen to test experimentally (**Figure 4.20**). Ligands were chosen with activation energies <25 kcal mol^{-1} for both transition states. Ligands were also chosen based on their commercial availability, if they were not commercially available the ligand must be easily synthesizable. Ligands L1-L8 were synthesised in-house, with L2-L4 being easily accessible analogues of L1 with varying electronic properties on the imine aryl group. Ligands L9-L15 are commercially available with L13 and L14 (1,10-phenanthroline and 2-isobutyrylcyclohexanone) being commonly used literature ligands for comparison.

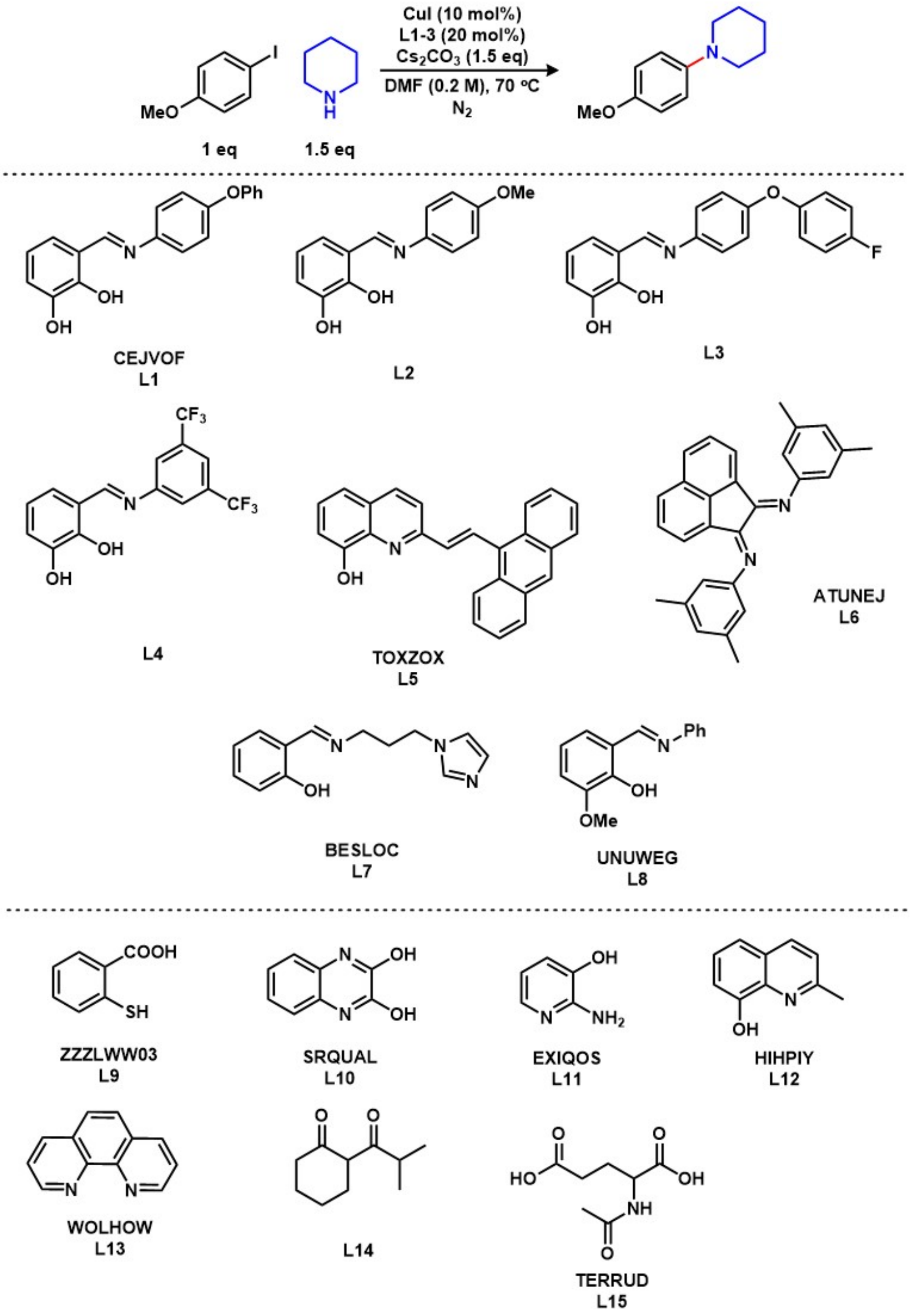


Figure 4.20: Structures of the ligands tested experimentally and their CSD identifiers.

All ligands were tested with the same set of reaction conditions as those used in the computational calculations. Piperidine was used as the amine coupling partner, caesium carbonate as the base, and DMF as the solvent. The aryl halide was changed from iodobenzene to 4-iodoanisole to enable easier analysis via NMR.

4.2.6.1 Activation Energies vs Yield

Activation energies were recalculated at 70° to match the experimental conditions. Calculated activation energies for both transition states along with experimental yields are shown in **Table 4.8**.

Table 4.8: Experimentally determined yields and calculated activation energies for the **TSOA** and **TSSig** transition states at both 25°C and 70°C.

Ligand	Average Yield (%)	Calculated Activation Energy TSOA at 25°C (kcal mol^{-1})	Calculated Activation Energy TSOA at 70°C (kcal mol^{-1})	Calculated Activation Energy TSSig at 25°C (kcal mol^{-1})	Calculated Activation Energy TSSig at 70°C (kcal mol^{-1})
L1	50	No TS	No TS	21.5	23.9
L2	14	No TS	No TS	22.2	24.5
L3	78	No TS	No TS	22.8	25.2
L4	67	22.5	24.8	22.1	24.6
L5	2	21.3	23.7	14.9	17.3
L6	1	18.6	21.0	17.5	19.9
L7	7	22.4	26.3	13.1	16.9
L8	3	20.5	25.4	12.3	17.2
L9	4	19.9	23.8	13.8	17.7
L10	1	22.5	24.8	22.1	21.3
L11	7	23.8	26.0	20.7	23.0
L12	7	23.0	25.3	19.3	21.7
L13	2	18.9	21.3	17.5	19.9
L14	100	19.9	22.2	15.8	18.1
L15	5	16.5	21.4	11.9	16.8

No trend is observed between calculated activation energies and experimental yield. Ligands with low activation energies, e.g. L6 and L15, have very low yields. Whereas ligands with high yields, L1-L4 and L14 have comparatively high activation energies. The same trend is observed across both transition states. Surprisingly L13, a commonly used literature ligand had a yield of only 2% in these reaction conditions. Ligands L1-L4 show moderate activity, these ligands have not been reported previously as ligands in the Ullmann-Goldberg reaction. The lack of identifiable oxidative addition transition states, even with enhanced criteria, suggests that the reaction proceeds via a sigma metathesis pathway for this class of ligands. However, radical mechanisms may be possible but were not explored. Experimental yield suggests that electron-withdrawing substituents on the imine increase the activity of this class of ligands, with L1, L3 and L4 showing reasonable yields with electron-withdrawing R groups. Using an electron-donating −PhOMe group, L2 shows a poor yield of only 14%. Successful identification of these ligands verifies that the data-mining approach is able to identify novel ligands in chemical databases. However, the interpretation of the output of the computational workflow is still not consistent as the activation energy does not correlate to experimental yield. This suggests that other aspects are contributing to yield apart from activation energy. This is consistent with the literature where deactivation pathways and ligand stability are key factors in catalyst activity. It may also be the case that the reaction is proceeding via a radical-based mechanism which was not explored in this work.

4.2.6.2 Deactivation Pathways

4.2.6.2.1 Disproportionation

$$2\,\mathrm{Cu}^{\mathrm{I}} \xrightarrow{\mathrm{DMF}} \mathrm{Cu}^{0} + \mathrm{Cu}^{\mathrm{II}} \tag{4.12}$$

Scheme 20: Disproportionation of copper(I) to copper(0) and copper(II) in DMF.

Copper(I) is known to disproportionate in solution to copper(0) and copper(II) deactivating the copper in the catalyst (**Scheme 20**). The Gibbs free energy of disproportionation was calculated at the B97-3c, TPSSh/def2-TZVP and DLPNO-CCSD(T)/def2-TZVPP levels of theory with DMF as the solvent using the SMD solvation model. Calculated energies are shown in **Table 4.9**.

Table 4.9: Gibbs free energies for the disproportionate of copper(I) to copper(0) and copper(II) in DMF. Gibbs free energies were calculated at the B97-3c, TPSSh/def2-TZVP and DLPNO-CCSD(T)/def2-TZVPP levels of theory.

Method	ΔG (kcal mol^{-1})
B97-3c	140
TPSSh/def2-TZVP	119
DLPNO-CCSD(T)/def2-TZVPP	124

All three methods suggest that the disproportionation of copper(I) to copper(0) and copper(II) in DMF is energetically unfavourable, suggesting that Cu(I) ions are stable in DMF and are not causing catalyst deactivation. However, this is a crude estimation as only an implicit solvent model was used which does not model solvent interactions, especially around ions, correctly. It also does not take into account the conversion of copper(0) from a solvated ion into a solid, which crashes out of solution.

4.2.6.2.2 Ligand Exchange

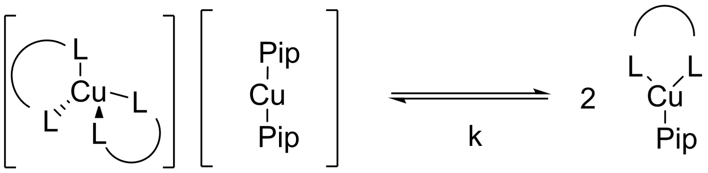


Figure 4.21: Equilibrium between the active catalytic species, **CuLpip** and inactive species, **[CuL$_2$][Cupip$_2$]**.

The main deactivation pathway, identified by Hartwig, was the equilibrium between the active catalyst state and a dimeric [CuL$_2$][CuNu$_2$] species.[182] The Gibbs free energy for the equilibrium (**Figure 4.21**) was calculated for all 15 experimentally tested ligands at the GFN2-xTB//B97-3c and TPSSh/def2-TZVP levels of theory in DMF. The results are shown in **Table 4.10**.

Table 4.10: Gibbs free energy of ligand exchange for the equilibrium between the active catalytic species, **CuLpip** and inactive species, **[CuL$_2$][Cupip$_2$]**. Gibbs free energies are calculated at the GFN2-xTB//B97-3c and TPSSh/def2-TZVP levels of theory.

Ligand	CSD_Refcode	Ligand Charge	Average Yield (%)	*k* (GFN2-xTB//B97-3c) (kcal mol^{-1})	*k* (TPSSh/def2-TZVP) (kcal mol^{-1})
L1	CEJVOF	−2	50	−7.6	−5.2
L2	-	−2	14	−8.0	−0.4
L3	-	−2	78	−7.7	2.8
L4	-	−2	67	−6.0	−4.9
L5	TOXZOX	−1	2	3.6	13.5
L6	ATUNEJ	0	1	−3.9	3.2
L7	BESLOC	−1	7	3.6	4.3
L8	UNUWEG	−1	3	7.5	3.1
L9	ZZZLWW03	−1	4	4.1	1.0
L10	SRQUAL	−1	1	−0.7	5.8
L11	EXIQOS	−1	7	−7.3	−6.8
L12	HIHPIY	−1	7	1.6	4.6
L13	WOLHOW	0	2	0.1	3.6
L14	L14	−1	100	0.1	7.3
L15	TERRUD	−1	5	20.4	−0.5

There is no clear trend between equilibrium energy and experimental yield for both GFN2-xTB//B97-3c and TPSSh/def2-TZVP. L1-L4 show negative values of k, which favour the formation of the inactive species. However, these ligands show good yields compared to the other ligands tested. **EXIQOS** has a similar value of k, however only has a 7% yield. L14, which shows the best yield of all the tested ligands only has a k value of 0.1 and 7.3 $\mathrm{kcal\,mol^{-1}}$ for B97-3c and TPSSh respectively. Ligands with a similar value of k, L5, L12 and L13, have poor yields in comparison. Ligands with large positive values of k, which favour the formation of the active catalytic species, also have poor yields experimentally. Ligand formal charge is also a poor predictor of activity. L1-L4 have a charge of -2 and show good yields compared to ligands with smaller formal charges. However, the best-performing ligand, L14, has a formal charge of -1. In comparison to other ligands with a -1 charge, L14 is an outlier with all other ligands performing poorly.

The poor correlation of calculated vs experimental results suggests that the experimental activity of ligands in the Ullmann-Goldberg reaction is a complex mixture of several different factors. This is supported by the lack of ligand understanding despite the amount of literature on mechanistic understanding and deactivation pathways for the reaction. This demonstrates the importance of mechanistic understanding in the guided design of homogeneous catalysis. The underlying chemical understanding is required to guide the targeted identification of new catalytic systems, otherwise, the chance of success is low. The availability of suitable tools to aid the design and identification of novel ligands and catalysts is important in streamlining the process of catalyst design.

4.3 Conclusions

CatSD and the developed high-throughput computational workflow were used to identify and predict the activity of ligands identified from the Cambridge Structural Database. The Ullmann-Goldberg reaction was used as a test reaction using both an amine and amide nucleophile. Over 10,000 ligands were identified from the CSD for both nucleophiles. Two mechanistic pathways, oxidative addition and sigma metathesis were explored. The high-throughput computational workflow showed excellent performance for the generation of complexes of interest and good performance for the calculation of transition states and activation energies. Success rates were limited by the general nature of the search and the unknown mechanistic pathway. CSD searches identified $\sim$ 12% unusable ligands, where both transition states failed.

Analysis of activation energies shows that for an amine nucleophile, both transition states are similar in energy. However, for an amide nucleophile, the sigma metathesis pathway is higher in energy. 15 ligands were identified and tested experimentally. Experimental validation identified a new class of ligands that have previously not been reported in the literature and the electronic effects of substituents for these ligands were explored. No direct link between activation energy and experimental yield was found. Disproportionation and deactivation energies were also explored with no direct link found between equilibrium energies and experimental yield.

In conclusion, the activity of ligands in the Ullmann-Goldberg reaction is likely a complex mixture of several different processes, involving mechanistic pathways, disproportionation and complex deactivation pathways. The understanding of chemical processes is paramount to successful ligand design for homogeneous catalysis.

4.4 Methodology

4.4.1 Experimental Methodology

All solvents and reagents were purchased from commercial sources. Solvents were HPLC standard and purchased from Sigma-Aldrich. Chemicals were purchased from Sigma-Aldrich (Dorset, UK) unless stated otherwise. Petroleum ether 40-60 °C was used unless stated otherwise. Nuclear Magnetic Resonance (NMR) spectra were recorded for ^{1}H at 400 MHz, ^{13}C at 101 MHz and ^{19}F at 376 MHz on a Bruker Ascend™ 400 spectrometer. Bruker Ascend™ 400 spectrometer was equipped with a multinuclear inverse probe for one-dimensional 1H and two-dimensional heteronuclear single quantum coherence (^{1}H$-$^{13}C HSQC), heteronuclear multiple bond correlation (^{1}H$-$^{13}C HMBC), and double quantum filtered correlation (^{1}H$-$^{1}H COSY). Chemical shifts (δ) are quoted in ppm downfield of tetramethylsilane. The coupling constants (J) are quoted in Hz (multiplicities: s singlet, bs broad singlet, d doublet, t triplet, q quartet and apparent multiplicities are described as m). Infrared (IR) spectra were recorded using a PerkinElmer Spectrum One FT-IR spectrophotometer or Bruker Alpha Platinum AR FTIR. Vibrational frequencies are reported in wavenumbers (cm^{-1}). Mass spectra (MS) were recorded on a maXis impact mass spectrometer employing electron spray ionization (ESI). All masses quoted are correct to four decimal places.

The analytical TLC chromatography was carried out using alumina-backed plates coated with silica gel 60 with a fluorescence indicator (20 x 20 cm, Merck) and visualised under ultra-violet (UV) radiation, $\lambda = 254$nm, and stained with potassium permanganate or ceric ammonium molybdate (CAM) followed by heating. Flash chromatography was carried out with silica (Geduran Si 60 with 40-63 µm) and sand.

Solvents were removed under reduced pressure using a Buchi rotary evaporator at 20 mbar, followed by further drying under high vacuum at 0.5 mmHg. To measure the melting points 2-3mm of the sample was placed in the bottom of the capillary tube and the tube was placed in the heating block. The 'melting point' was measured as the range from the appearance of the first liquid droplet until the complete melting of the crystals. The measurements were carried out on a Mettler Toledo melting point system.

4.4.1.1 Ligand Synthesis

Ligand L9-L15 were purchased from commercial sources.

4.4.1.1.1 Preparation of L1

2,3-dihydroxybenzaldehyde (10.00 mmol, 1.38 g) and 4-phenoxyaniline (9.99 mmol, 1.85 g) was added to ethanol (100 mL) and refluxed under nitrogen for 5 h. The solution was cooled to room temperature and concentrated in vacuo. The product was separated by filtration, recrystallized in ethanol then dried under vacuum to afford **L1** (2.48 g, 81.3%), as dark red crystals.

δH (400 MHz, $CDCl_3$): 8.63 (1H, s), 7.40 (2H, ddd, J = 2.28, 4.28, 9.87 Hz), 7.32 (2H, dt, J = 3.19, 8.80 Hz), 7.17 (1H, t, J = 7.4 Hz), 7.11-7.06 (5H, m), 6.98 (1H, dd, J = 1.39, 7.83 Hz), 6.86 (1H, t, J = 7.85 Hz).

δC (101 MHz, $CDCl_3$): 161.27, 156.97, 156.60, 149.35, 145.15, 142.80, 129.91, 123.65, 122.89, 122.41, 119.62, 119.04, 118.52, 117.53. mp 144.5-145.9 °C.

4.4.1.1.2 Preparation of L2

2,3-dihydroxybenzaldehyde (9.99 mmol, 1.38 g) and *p*-anisidine (9.98 mmol, 1.23 g) was added to methanol (60 mL) and refluxed under nitrogen for 3 h. The solution was allowed to cool to room temperature for 48 h. The product was separated by filtration and dried under vacuum to afford **L2** (2.18 g, 89.9%), as dark red crystals.

δH (400 MHz, $CDCl_3$): 8.63 (1H, s), 7.31 (2H, dt, J = 3.38, 8.95 Hz), 7.05 (1H, dd, J = 1.43, 7.84 Hz), 6.97 (3H, dt, J = 3.35, 8.89 Hz), 6.84 (1H, t, J= 7.86 Hz), 3.87 (3H, s).

δC (101 MHz, $CDCl_3$): 160.14, 159.11, 149.82, 145.31, 140.44, 122.79, 122.29, 118.92, 118.60, 117.28, 114.84, 55.69. mp: 131.4-132.5 °C.

4.4.1.1.3 Preparation of L4

2,3-dihydroxybenzaldehyde (9.99 mmol, 1.38 g) and 4-fluoroaniline (0.95 mL, 10.00 mmol) was added to methanol (60 mL) and refluxed under nitrogen for 5 h. The solution was allowed to cool to room temperature for 48 h. The product was separated by filtration and dried under vacuum to afford **L3** (2.77 g, 94.1%), as red-orange crystals.

δH (400 MHz, DMSO): 13.03 (1H, s), 9.22 (1H, s), 8.90 (1H, s), 7.46-7.51 (2H, m), 7.32-7.27 (2H, m), 7.09 (1H, d, J = 7.56 Hz), 6.96 (1H, d, J = 7.32 Hz), 6.82 (1H, t, J = 7.77 Hz).

δC (101 MHz, $CDCl_3$): 163.01, 162.22, 160.53, 148.90, 145.05, 123.01, 122.58, 119.22, 118.49, 117.79, 116.34.

δF (376 MHz, $CDCl_3$): -115.15 (q, J = 3.76 Hz). v_{max} (neat): 3451, 1658, 1461, 1273. mp: 107.3-108.5 °C, m/z (HRMS) Found MS^+= 232.0757.

4.4.1.1.4 Preparation of L4

2,3-dihydroxybenzaldehyde (9.43 mmol, 1.30 g) and 3,5-bis(trifluoromethyl)aniline (1.47 mL, 9.43 mmol) was added to ethanol (100 mL) and refluxed under nitrogen for 2 h. The solution was allowed to cool to room temperature and concentrated in vacuo. The product was recrystallised in chloroform:n-hexane = 3:2, separated by filtration and dried under vacuum to afford **L4** (1.12 g, 33.9%), as dark red crystals.

δH (400 MHz, DMSO$-d_6$): 12.20 (1H, s), 9.35 (1H, s), 9.05 (1H, s), 8.13 (2H, s), 7.98 (1H, s), 7.17 (1H, dd, J = 1.29, 7.71 Hz), 7.01 (1H, dd, J = 1.44, 7.82 Hz), 6.83 (1H, t, J = 7.79 Hz).

δC (101 MHz, $CDCl_3$): 167.77, 151.10, 149.64, 146.20, 131.84, 127.75, 125.03, 123.45, 122.30, 120.24, 119.9, 119.56.

δF (376 MHz, $CDCl_3$): -61.1. mp 113.7-116.4 °C.

4.4.1.1.5 Preparation of L5

2-methyl-8-quinolinol (10.01 mmol, 1.59 g) and 9-anthracenecarboxaldehyde (10.99 mmol, 2.27 g) was added to a dried 25 mL round bottom flask and purged with nitrogen for 10 mins. Acetic anhydride (10 mL) was added and refluxed under nitrogen for 20 h. The solution was cooled to room temperature, poured onto ice water (200 mL), and stirred overnight. The slurry was filtered and washed with water. The resulting powder was dissolved in DCM (50 mL), washed with water (2 x 50 mL) and dried with $MgSO_4$. The solution was concentrated in vacuo and purified using flash column chromatography (DCM:pentane = 1:3). The product was recrystallised in DCM and dried under vacuum to afford **L5** (0.74 g, 21.3%) as yellow crystals.

δH (400 MHz, $CDCl_3$): 8.56 (1H, dd, J = 0.6, 16.44 Hz), 8.40 (1H, s), 8.36-8.31 (2H, m), 8.15 (1H, d, J = 8.48 Hz), 8.00-7.96 (2H, m), 7.71 (1H, d, J = 8.48 Hz), 7.46-7.42 (4H, m), 7.39 (1H, t, J = 7.90 Hz), 7.30 (1H, dd, J = 1.23, 8.27 Hz), 7.19-7.12 (3H, m).

δC (101 MHz, $CDCl_3$): 153.29, 152.21, 138.11, 136.72, 136.69, 131.69, 131.50, 131.07, 129.70, 128.85, 127.75, 127.57, 127.26, 125.90, 125.79, 125.33, 120.62, 117.77, 110.30. mp 208.3- 208.5 °C.

4.4.1.1.6 Preparation of L6

Acenaphthenequinone (1.00 g, 5.49 mmol) and zinc chloride (2.41 g, 17.71 mmol) were stirred in glacial acetic acid (10 mL). The solution was heated to 60 °C and 3,5-dimethylaniline (1.49 mL, 12.63 mmol) was added and the mixture refluxed for 1 h. The solution was filtered while hot and washed with water and diethyl ether. The BIAN zinc complex was dissolved in DCM (100 mL) and was added to a separating funnel. Sodium oxalate (2.59 g, 19.36 mmol) in water (100 mL) was added and shaken for 5 mins. The organic layer was separated, washed with water (x2), and dried with $MgSO_4$. The slurry was filtered and evaporated to dryness to yield give the crude product.

δH (400 MHz, $CDCl_3$): 8.18 (1H, d, J = 8.26 Hz), 8.01 (1H, d, J = 7.11 Hz), 7.80-7.74 (3H, m), 7.29 (2H, t, J = 7.57 Hz), 6.82-6.65 (8H, m), 2.28 (12H, s).

δC (101 MHz, $CDCl_3$): 161.05, 151.94-115.69, 21.54. mp 219.1-221.6 °C.

4.4.1.1.7 Preparation of L7

2-hydroxybenzaldehyde (1.71 mL, 16.4 mmol) was stirred with 1-(3-aminopropyl)imidazole (1.95 ml, 16.4 mmol) in ethanol (20 mL). The solution was refluxed for 2 h, cooled to room temperature and concentrated in vacuo. The resulting solid was washed with petroleum ether (40-60) and dried under vacuum to afford the crude product.

δH (400 MHz, $CDCl_3$): 13.07 (1H, s), 8.25 (1H, s), 7.39 (1H, s), 7.26 (1H, td, J = 1.74, 5.59 Hz), 7.21-7.15 (1H, m), 7.01 (1H, t, J = 1.03 Hz), 6.91 (1H, d, J = 8.20 Hz), 6.88-6.80 (1H, m), 4.00 (2H, t, J = 6.89 Hz), 3.49 (2H, td, J = 1.14, 5.24 Hz), 2.13 (2H, p, J = 6.72 Hz).

δC (101 MHz, $CDCl_3$): 166.13, 160.92, 137.14, 132.60, 131.44, 129.84, 118.88, 118.72, 118.57, 117.01, 55.89, 44.32, 31.80. mp 86.9-87.4 °C.

4.4.1.1.8 Preparation of L8

o-vanillin (1.52 g, 10.01 mmol) and aniline (0.91 mL, 10.00 mmol) was dissolved in methanol (50 mL) and refluxed for 1 h. The solvent was removed under vacuum, 10 drops of petroleum ether (40-60) was added and left to stand overnight. The product was dried under vacuum to yield the crude product as a yellow powder.

δH (400 MHz, DMSO$-d_6$): 13.25 (1H, s), 8.96 (1H, s), 7.51- 7.45 (2H, m), 7.44-7.40 (2H, m), 7.35-7.30 (1H, m), 7.25 (1H, dd, J = 1.48, 7.86 Hz), 7.14 (1H, dd, J = 1.48, 8.05 Hz), 6.92 (1H, t, J = 7.93 Hz), 3.83 (3H, s).

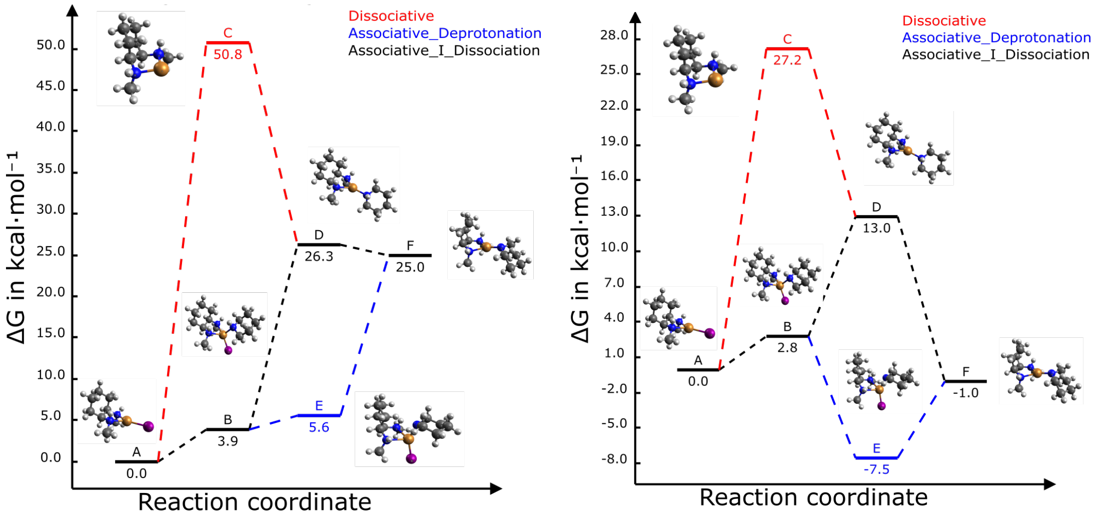
δC (101 MHz, DMSO$-d_6$): 164.15, 151.25, 148.39, 148.36, 129.95, 127.46, 124.40, 121.82, 119.67, 119.07, 116.07, 56.35. mp 81.2-81.6 °C.

4.4.1.2 Ligand Testing

In a nitrogen-filled glove box, a stock solution was prepared containing 4-iodoanisole (1.3 mg, 5.56 mmol), piperidine (824 µL, 8.32 mmol), 1,2,4,5-tetramethylbenzene (internal standard, 187 mg, 1.39 mmol) and DMF (27.8 mL). These stock solutions were transferred (2.14 mL each) to 12 separate vials which were charged with CuI (8.2 mg, 0.0428 mmol), Cs_2CO_3 (208 mg, 0.640 mmol) and ligand (0.0856 mmol). The resulting reaction mixtures were taken out of the glove box and heated at 70 °C for 16 h. After this time, the reaction mixtures were exposed to air, diluted with EtOAc (20 mL), washed with H_2O (5 x 5 mL), dried (Na_2SO_4), filtered and concentrated under reduced pressure. The resulting residue was analysed by 1H NMR ($CDCl_3$). Each ligand was tested 2 times and an average yield taken.

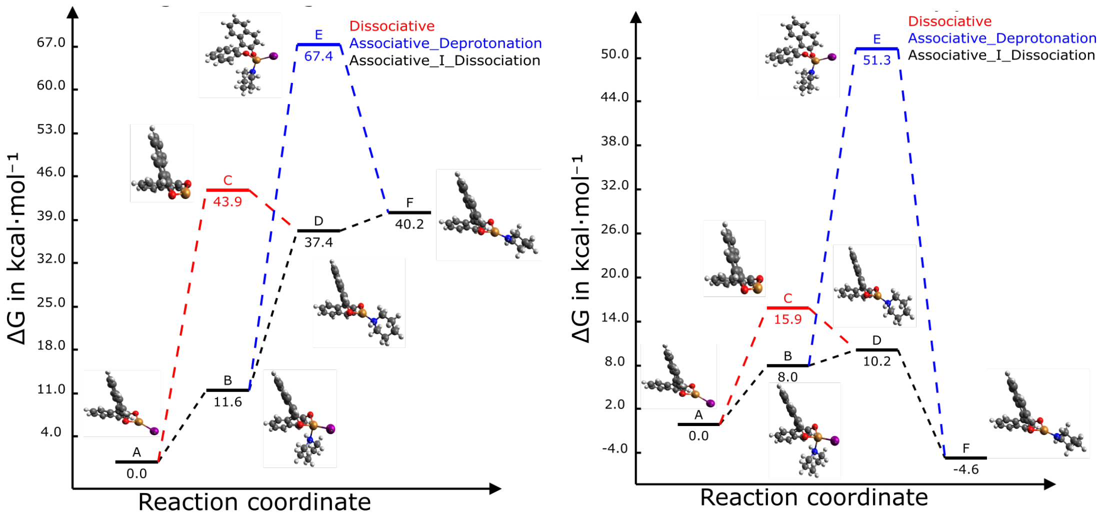
4.A Appendix

4.A.1 Ligand Exchange



(a) GFN2-xTB//B97-3c **(b)** DLPNO-CCSD(T)/def2-TZVPP

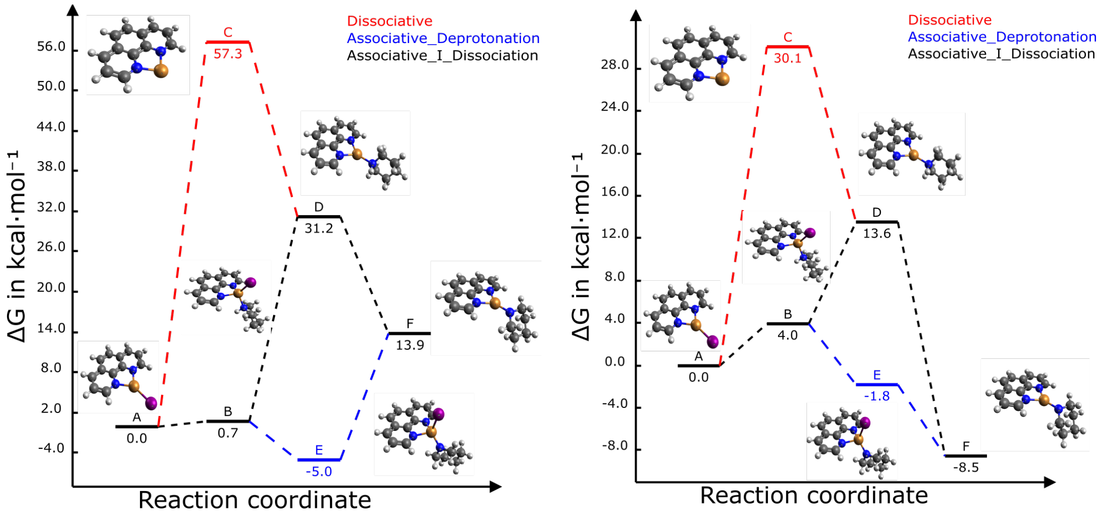
Figure 4.A.1: Ligand exchange mechanisms for ligand **L0011** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c

(b) DLPNO-CCSD(T)/def2-TZVPP

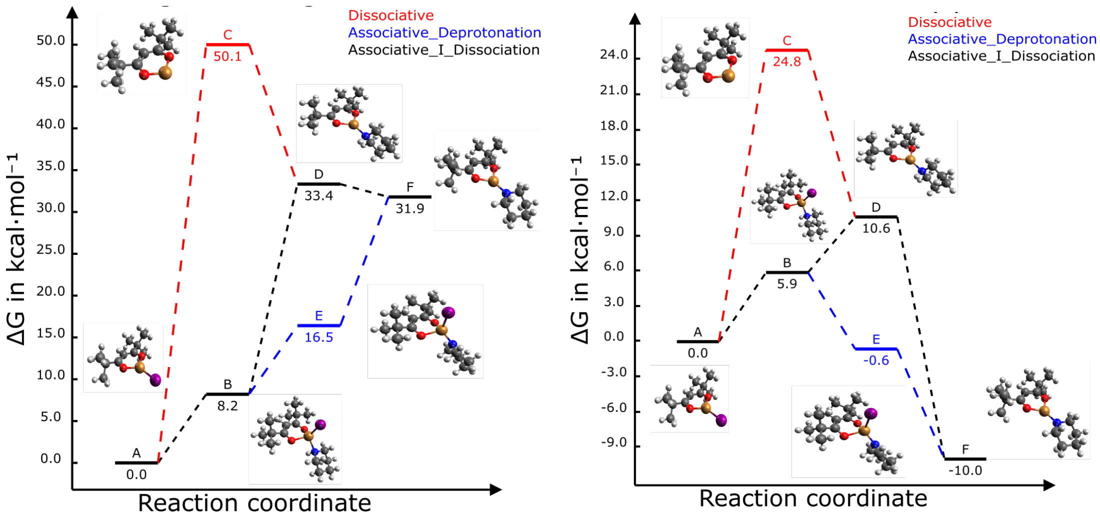
Figure 4.A.2: Ligand exchange mechanisms for ligand **L0039** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c

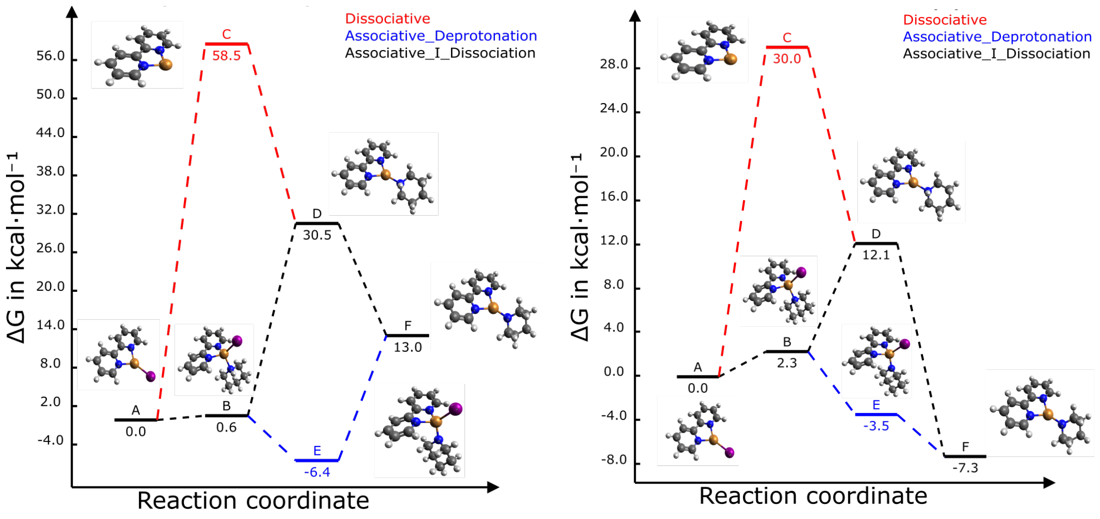
(b) DLPNO-CCSD(T)/def2-TZVPP

Figure 4.A.3: Ligand exchange mechanisms for ligand **L0042** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



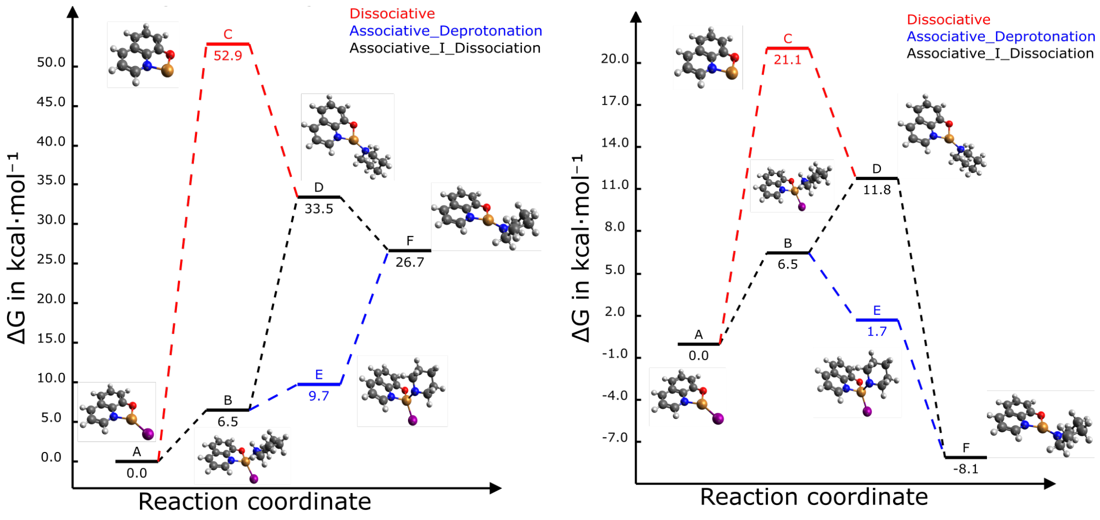
(a) GFN2-xTB//B97-3c **(b)** DLPNO-CCSD(T)/def2-TZVPP

Figure 4.A.4: Ligand exchange mechanisms for ligand **L0101** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c **(b)** DLPNO-CCSD(T)/def2-TZVPP

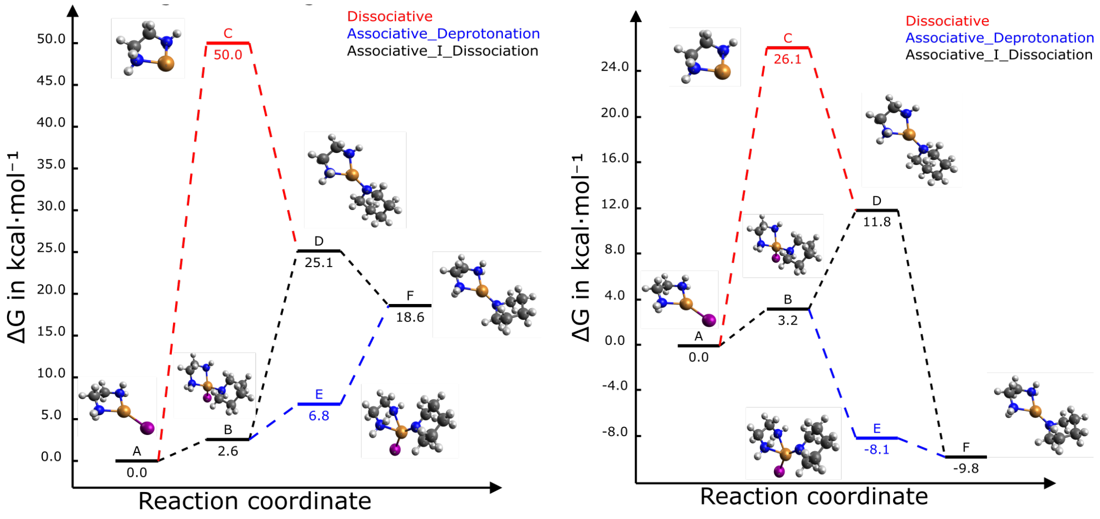
Figure 4.A.5: Ligand exchange mechanisms for ligand **L0104** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c

(b) DLPNO-CCSD(T)/def2-TZVPP

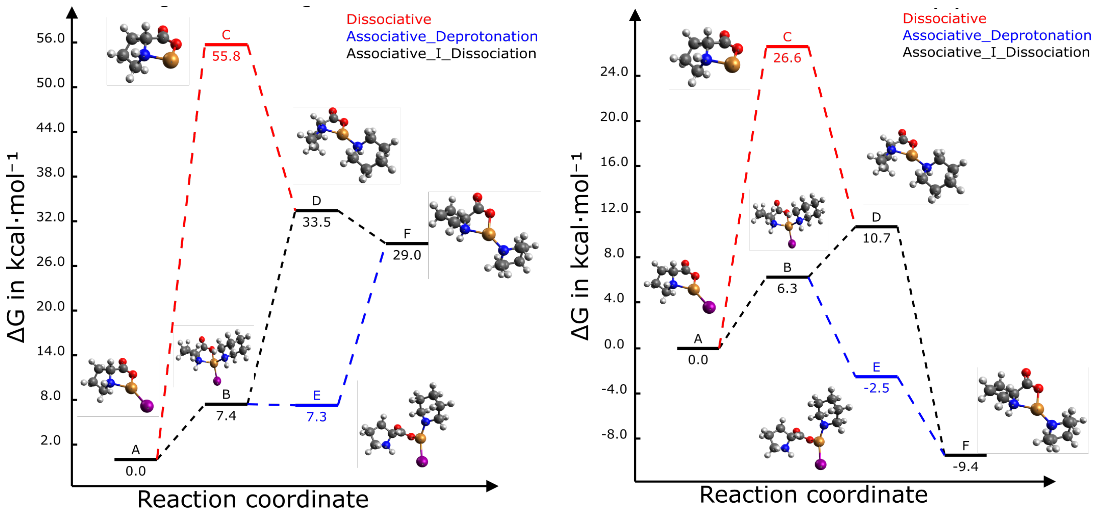
Figure 4.A.6: Ligand exchange mechanisms for ligand **L0144** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c

(b) DLPNO-CCSD(T)/def2-TZVPP

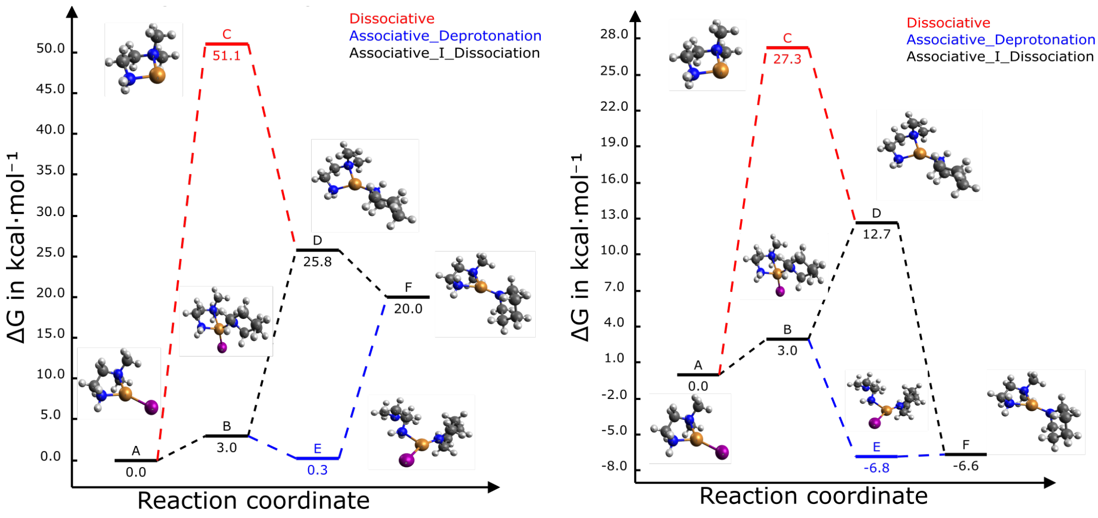
Figure 4.A.7: Ligand exchange mechanisms for ligand **L0177** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c

(b) DLPNO-CCSD(T)/def2-TZVPP

Figure 4.A.8: Ligand exchange mechanisms for ligand **L0197** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.



(a) GFN2-xTB//B97-3c

(b) DLPNO-CCSD(T)/def2-TZVPP

Figure 4.A.9: Ligand exchange mechanisms for ligand **L0235** from the ***ligands_lit_set*** calculated at the GFN2-xTB//B97-3c and DLPNO-CCSD(T)/def2-TZVPP levels of theory.

4.A.2 Literature Ligand Analysis

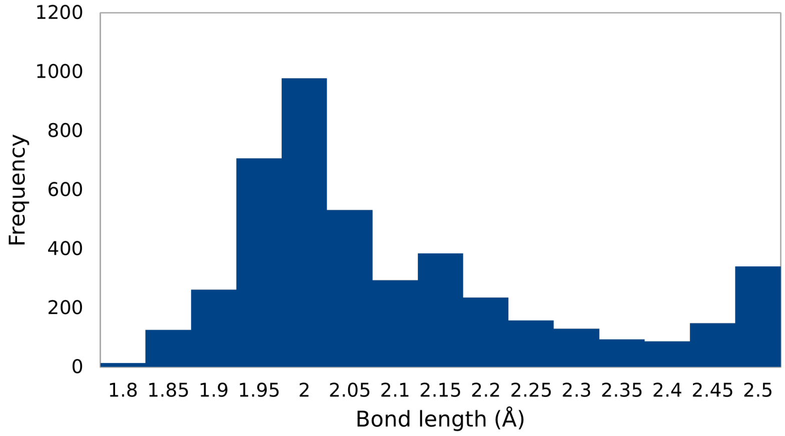


Figure 4.A.10: Distribution of Cu−L bond lengths in ***ligands_lit_set***.

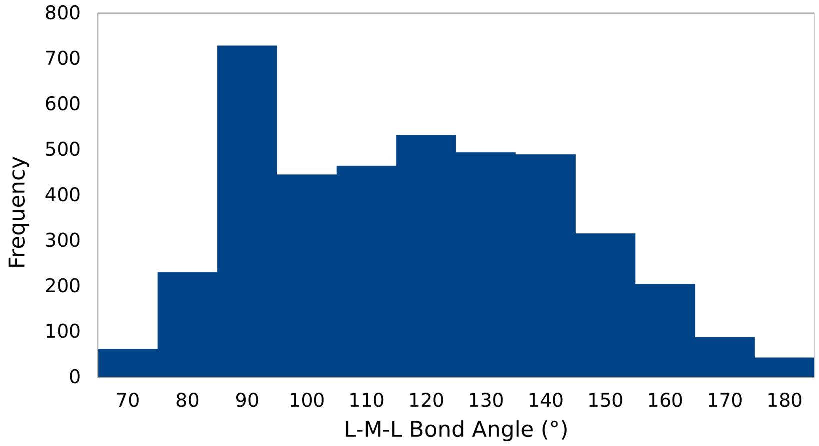
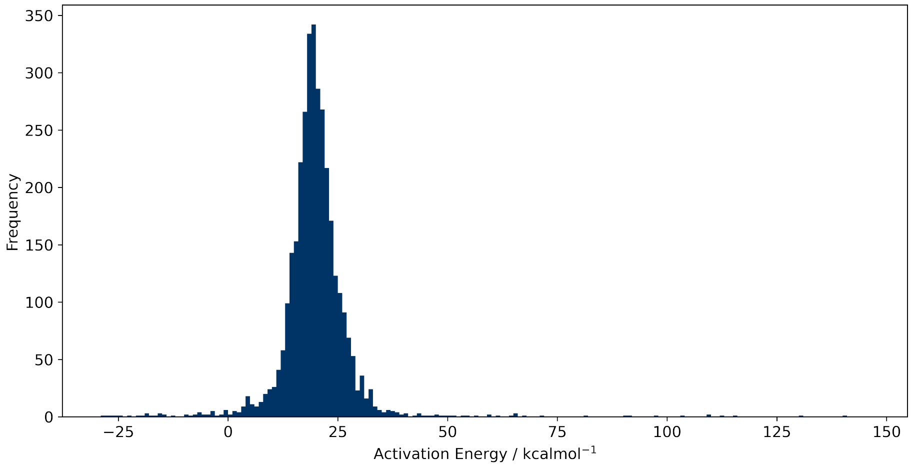
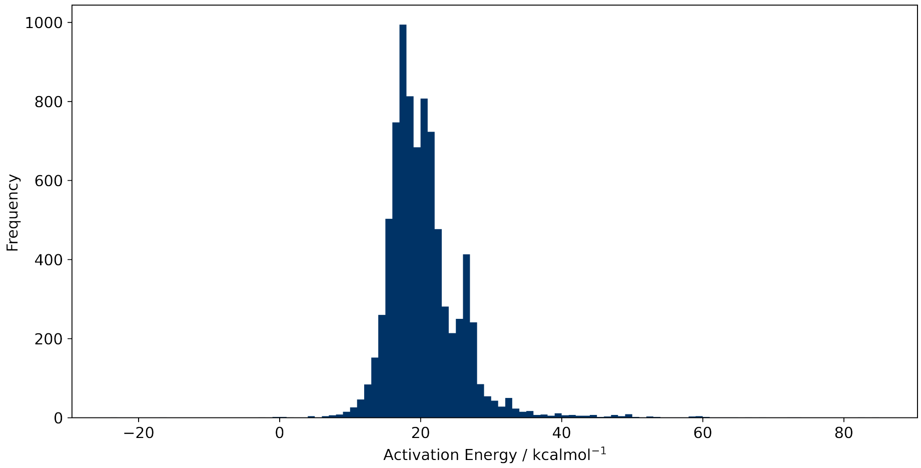


Figure 4.A.11: Distribution of L−cu−L bond angles in ***ligand_lit_set***

4.A.3 Activation Energy Analysis

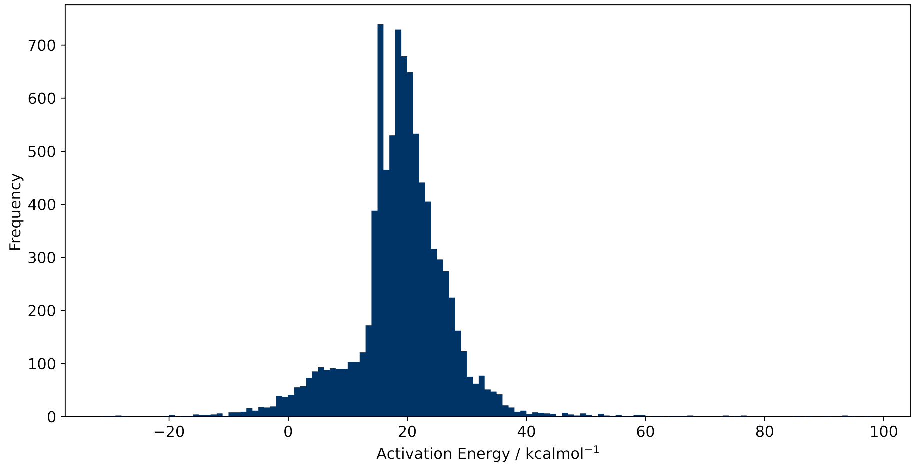


(a) TSOA

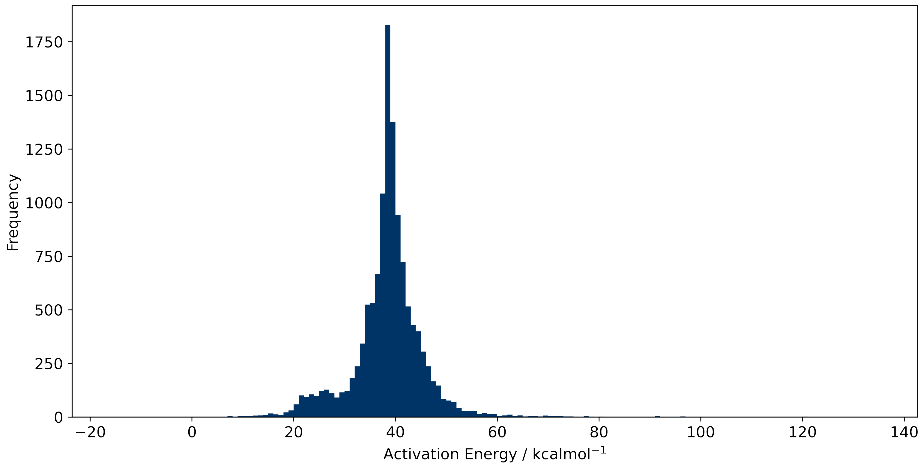


(b) TSSig

Figure 4.A.12: Distribution of activation energies for both the **TSOA** and **TSSig** transition states for piperidine. Full range of activation energies.



(a) TSOA



(b) TSSig

Figure 4.A.13: Distribution of activation energies for both the **TSOA** and **TSSig** transition states for 2-pyrrolidinone. Full range of activation energies.

4.A.4 Experimental Validation

Table 4.A.1: Individual yields for both experimental tests with each ligand used to take the average yield.

Ligand	Yield 1 (%)	Yield 2 (%)	Average Yield (%)
L1	50	49	50
L2	12	16	14
L3	88	68	78
L4	78	56	67
L5	2	2	2
L6	1	1	1
L7	7	7	7
L8	3	3	3
L9	3	4	4
L10	0	2	1
L11	6	8	7
L12	7	6	7
L13	2	2	2
L14	101	99	100
L15	6	4	5

4.A.5 Ligand Exchange

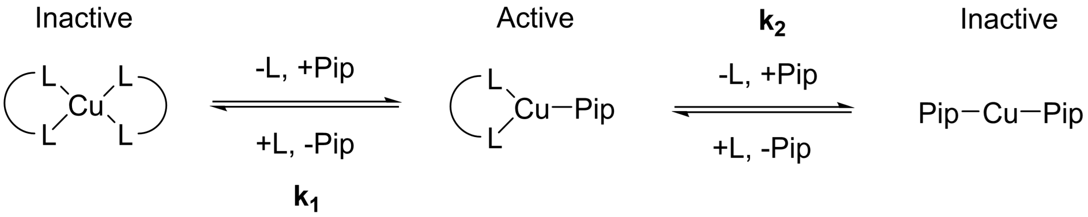


Figure 4.A.14: Equilibrium between the active catalytic species, **CuLpip** and inactive species, **CuL$_2$** and **Cupip$_2$**.

Table 4.A.2: Gibbs free energy of ligand exchange for the equilibrium between the active catalytic species, **CuLpip** and inactive species, **CuL$_2$** and **Cupip$_2$**. Gibbs free energies are calculated at the GFN2-xTB//B97-3c level of theory.

Ligand	CSD_Refcode	Average Yield (%)	k_1 (kcal mol^{-1})	k_2 (kcal mol^{-1})
L1	CEJVOF	50	32.9	−40.5
L2	DAHLEG_Alt	14	31.7	−39.7
L3	L3	78	32.7	−40.4
L4	HAGXUJ_Alt	67	37.3	−43.3
L5	TOXZOX	2	28.2	−24.6
L6	ATUNEJ	1	−9.3	5.3
L7	BESLOC	7	20.4	−16.7
L8	UNUWEG	3	23.5	−16.0
L9	ZZZLWW03	4	27.5	−23.4
L10	SRQUAL	1	−21.3	20.6
L11	EXIQOS	7	16.4	−23.7
L12	HIHPIY	7	79.0	−77.4
L13	WOLHOW	2	−9.3	9.4
L14	L14	100	−18.4	18.5
L15	TERRUD	5	28.3	−7.9

Chapter 5: Machine Learning Prediction of Activation Energies

5.1 Introduction

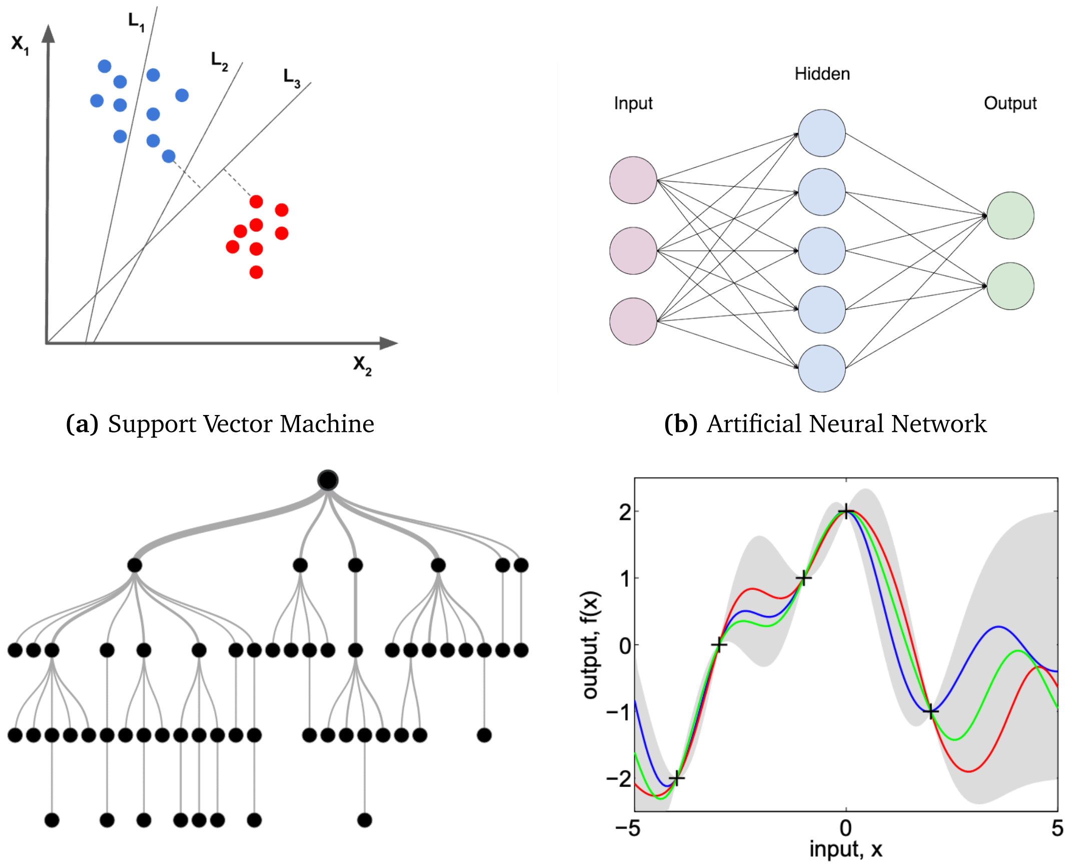
5.1.1 Machine Learning Methods

Machine learning is a family of methods for automated data analysis, capable of detecting complex patterns in data and illustrating the intricate connection between descriptors and desired output. Regression-based machine learning algorithms build a model using available data and the output is estimated through quantitatively learning from adequate training samples.

The full range of machine learning methods have been applied to chemical problems. Simple regression models such as Multiple Linear Regression (MLR) tend to underfit data due to their simplicity and inability to use non-linear data. MLR models are simple, fast to train and easy to interpret providing an initial indication of the applicability of machine learning to a specific application. Partial Least Squares (PLS) can compress the data into fewer variables so are often used for noisy datasets containing many descriptors.[209]

Support Vector Machines (SVM) and tree-based models such as Random Forest and ExtraTrees are robust and versatile machine learning methods which have been applied to a range of problems in chemistry.[210] Random forests have been shown to perform well on transition metal complexes for the prediction of spin-splitting, metal-ligand bond lengths and redox potentials.[52]

Artificial Neural Networks have recently found wide application in the computational chemistry community.[211] Broader applications have been present recently in materials science, molecular and heterogeneous catalysis.[212,213] ANNs are the most commonly used machine learning model used in transition metal applications and have been used both with simple regression and in conjunction with molecular graphs to predict complex transition metal quantum properties, spin-splitting and metal-ligand bond lengths with errors of 3 $\text{kcal}\,\text{mol}^{-1}$ and 0.02-0.03 Å respectively.[51,53] ANNs tend to overfit (fitting the data too tightly) and require the determination of a suitable set of broadly applicable descriptors that enable the use of ANNs beyond the molecules in the training set (e.g. for larger molecules and more diverse chemistry). ANNs are extremely adaptable and have shown to be effective in computational and organic chemistry however, their applicability to transition metal-based problems is extremely dependent on the development of a suitable descriptor set, especially for open-shell transition metal complexes. ANNs have shown to be difficult to use for transition metal complexes for the following reasons: (i) the process of designing the ANN and its architecture is more involved than other machine learning methods, (ii) overfitting can be a problem, especially in noisy datasets, (iii) they generally do not perform well for small datasets, (iv) development of a suitable descriptor set is complex and (v) it is more difficult to create interpretable models, especially for models where a large number of descriptors are required.



(a) Support Vector Machine

(b) Artificial Neural Network

(c) Tree-based models

(d) Gaussian Process Regression

Figure 5.1: Types of regression machine learning models.

Finally, Gaussian Process Regression (GP) is becoming more popular in chemistry, but its uptake in transition metal-containing applications is low with very few examples available.[214] GP has a similar fitting process to SVM but also provides a prediction distribution making uncertainties in individual prediction easy to calculate.

In summary, MLR and PLS are non-linear methods which are expected to perform poorly in transition metal problems. ANN is an extremely powerful method but is difficult to use due to the requirement for large datasets and requires a carefully developed descriptor set to be effective. There are very few examples of GP being applied to transition metal systems, however, this method offers the advantage of error bars and robust fitting. SVM and tree-based methods are widely applicable methods that should prove effective for this application.

5.1.2 Descriptors used in Organometallic Catalysis

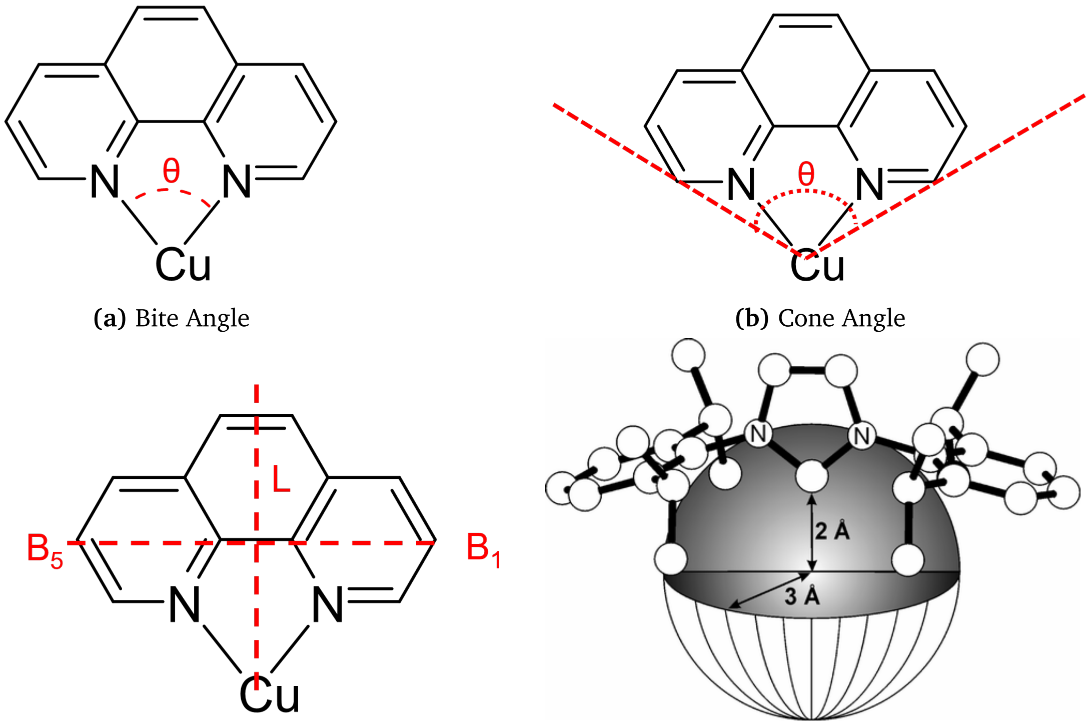
Descriptors that work well for organic molecules have proven to be unsuitable for inorganic molecules.[51] The electronic properties of transition metal complexes (e.g. spin state splitting) are incredibly sensitive to the ligand coordinating atom identity which dominates the ligand field strength.[53] More sophisticated descriptors need to be used to effectively capture transition metal bonding, due to its dependence on coordination environment.[215]

Bidentate ligands provide a more well-defined coordination environment than monodentate ligands due to occupying two (generally *cis*) sites on the transition metal centre. Despite their frequent use synthetically, the characterisation of bidentate ligands through calculated ligand descriptors is less commonplace than for monodentate ligands. Most available software for calculating organometallic ligand descriptors only apply to monodentate ligands.[216]

5.1.2.1 Steric Descriptors

The simplest type of steric descriptors are bond lengths and bond angles. Both descriptors can be considered net interaction descriptors as they describe both steric and electronic properties. Bond lengths are a measure of both the distance between two atoms, (e.g. the distance of the ligand from the metal centre) and the strength of the bond. Bond angles describe both the angle occupied by a set of atoms and the amount of distortion present at an atom. Bond length and angles can be calculated from atomic coordinates and are usually generated for atoms of interest, e.g. ligand coordinating atoms and transition state active atoms.

The oldest and most common steric descriptor is the ligand bite angle ($\angle$L$-$M$-$L) which measures the net interaction with the metal centre, capturing a combination of both steric and electronic effects.[217] Bite angle provides an easily accessible measure of sterics, which can be easily calculated from a set of atomic coordinates. Cone angle is another steric descriptor which describes the angle of a cone, originating from the metal centre, which encompasses the entire ligand. Cone angle was included in Tolman's 1977 review but has not been adopted widely for bidentate ligands.[34] Allen and coworkers proposed an improved, mathematically rigorous method to determine an exact cone angle (θ°) by solving for the most acute right circular cone that contains the entire ligand.[218] The procedure is applicable to any ligand, planar or nonplanar, monodentate or polydentate, bound to any metal centre in any environment.[218]



(a) Bite Angle

(b) Cone Angle

(c) Sterimol Parameters

(d) % Buried Volume[219]

Figure 5.2: Commonly used steric descriptors for bidentate organometallic ligands.

Buried volume ($\%V_{Bur}$), gives a measure of the volume occupied by the ligand in the first coordination sphere of the metal centre, or in a sphere of radius r.[219] Buried volume was originally created to describe NHC ligands but has since been applied to P and N donor ligands.[21,220] The Sterimol parameters were developed by Verloop to describe the steric size of substituents.[221] The Sterimol parameters were originally developed for organic compounds but three of the five parameters, L, B_1 and B_5, have been applied to organometallic ligands.[221] The metal atom is called atom 1 and the first atom in the substituent is called atom 2. L can be described as the depth of the substituent. It is defined as the length of the vector going from atom 1, through atom 2 and ending on the tangent of the vdW surface. B_1 and B_5 can be described as the minimum and maximum rotational size of the substituent. They are defined as the shortest and longest vectors at a tangent from L to the vdW surface, respectively. Solvent-accessible surface area is a measure of how much of the area of a molecule is available to the solvent. The atomic SASA can be used as a measure of the steric availability of an individual atom.

The He_8 descriptor was developed by Purdie et al. to describe the steric bulk of P,P donor ligands in an octahedral coordination environment.[36] He_8, is calculated as the interaction energy between a ligand in its chelating conformation (from zinc complex, $[ZnCl_2(LL)]$) and a rigid wedge of eight helium atoms.

5.1.2.2 Electronic Descriptors

The majority of electronic descriptors used to describe both mono- and bidentate ligands require the generation of a new complex with a standard set of auxiliary ligands to allow for standardisation and comparison between ligands. These descriptors measure the CO bond strength in a standard complex to determine the strength of metal-ligand interaction. Increasing the number of these descriptors dramatically increases the computational time required to calculate descriptor sets. For example, Tolman's electronic parameter (TEP), and its derivatives (CEP, SEP, LTEP), require a $Ni(CO)_3L$ tetrahedral complex to be generated and then the stretching frequency of the CO bond calculated.[34] Other electronic descriptors found in ligand knowledge bases, such as bond dissociation energy and proton affinity also require the generation of addition complexes to derive the value. Natural bonding orbital charges provide a measure of electron density at specific atoms but require additional software to use. Properties such as the energy of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) are much easier to calculate and can be taken directly from a computational chemistry output file. Population analysis is an alternative method of determining the electronic properties of specific atoms in a complex.

Population analysis is the determination of the distribution of electrons in a molecule. It allows the direct description of chemical properties for any given atom of interest and can be obtained directly from standard computational output files. Several methods for performing population analysis have been developed, Mulliken, Löwdin and Mayer.[222–224] Each method analyses the SCF wavefunction of a computational chemistry calculation to generate a large set of molecular and atomic properties. Properties include atomic and orbital charges, bond orders and bonded valence. Mulliken population analysis is very sensitive to the basis set used. Electrons are assigned to a single atom based on the basis functions of the entire molecule. This method has no basis set limit and therefore the exact value is dependent on the way the limit is approached, meaning charges are ill-defined and have no exact value. Swapping basis sets may yield completely different results for the same compound. Mulliken population analysis also partitions electronegativity in such a way that charges separations in a molecule can become exaggerated. Löwdin's approach avoids the problem of negative populations or populations greater than 2. Löwdin charges are often closer to chemically intuitive values and are less sensitive to the basis set. In Mayer's population analysis, the charges are identical to that of Mulliken's, however, it provides much better values for bond orders and bonded valences.

Descriptors based on population analysis have been used previously for the prediction of C−N cross-coupling with palladium.[57] Single-atom electronic descriptors derived from population analysis provide a targeted description of key properties such as charge, electron density and bond strength for specific atoms of interest, with little to no additional computational cost. Descriptors can be quickly and easily extracted from computational output files using Python without the need for additional treatment.

5.2 Results and Discussion

5.2.1 Descriptor Selection and Extraction

In order to describe the properties of the computationally generated transition states a set of chemically relevant descriptors was designed. The set of descriptors must be obtainable through the output of the high-throughput computationally workflow, either from the energy and frequency calculations or calculated from the 3D optimised structures. A good descriptor set should be cheap to compute, contain as few dimensions as possible and preserve target similarity (i.e. complexes with similar properties should have similar feature representations). Descriptors were chosen to cover both the steric and electronic properties of the transition state. Steric descriptors were chosen to describe to steric properties of the ligand around the copper centre. To ensure completeness steric descriptors were chosen to ensure that all steric properties were described, including lengths, angles, volume and area.

Bond lengths and bond angles were calculated straight from the 3D structure output from the geometry optimisation. Sterimol parameters, Cone Angle (Allen and coworkers' exact cone angle), and Solvent Accessible Surface Area (SASA) were calculated using the morfeus python module.[216] The morfeus package calculates properties for monodentate ligands by default. For bidentate ligands, the structure has to be adjusted to ensure the calculation correctly describes the ligand. To ensure Cone Angle and Sterimol parameters (B1, B5 and L) are calculated correctly a dummy hydrogen atom is placed at the midpoint of the two ligand coordinating atoms. The indexes of the two coordinating atoms, L1 and L2, define the xy and z planes. For Buried Volume, the substrates were removed from the complex to ensure that only the Buried Volume of the ligand is calculated. Buried Volumes were calculated at 3.5 Å, 5 Å and 7 Å with hydrogen atoms, excluding the dummy atom, included. For descriptors describing a change in length or angle, the difference is taken between the transition state and the **CuLI** intermediate. An overview of the structure pre-treatment is shown in **Figure 5.3**.

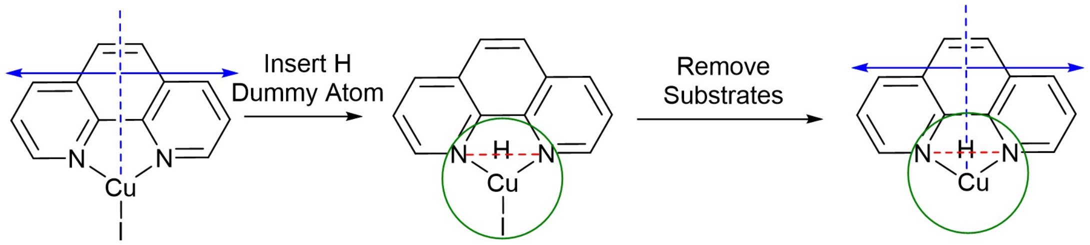


Figure 5.3: Treatment of the 3D structure of each complex for calculation of steric descriptors. Sterimol (blue) and % Buried Volume (green).

Electronic descriptors were chosen to model the electronic properties of the key atoms; the copper centre, ligand coordinating atoms, nucleophile nitrogen and the carbon and iodine atoms in the aryl halide. Several electronic properties were encapsulated in these descriptors; atomic and orbital charges, electron population, bond strength, valence and the potential energy surface. All descriptors were extracted from the energy and frequency output files from ORCA using a customised version of the cclib library to add functionality for the extraction of HOMO and LUMO energies, Mayer's bond order, Mayer's bonded valence, Mulliken atomic population, Löwdin atomic charges and Löwdin orbital charges.[225] A full list of descriptors is shown in **Table 5.1**.

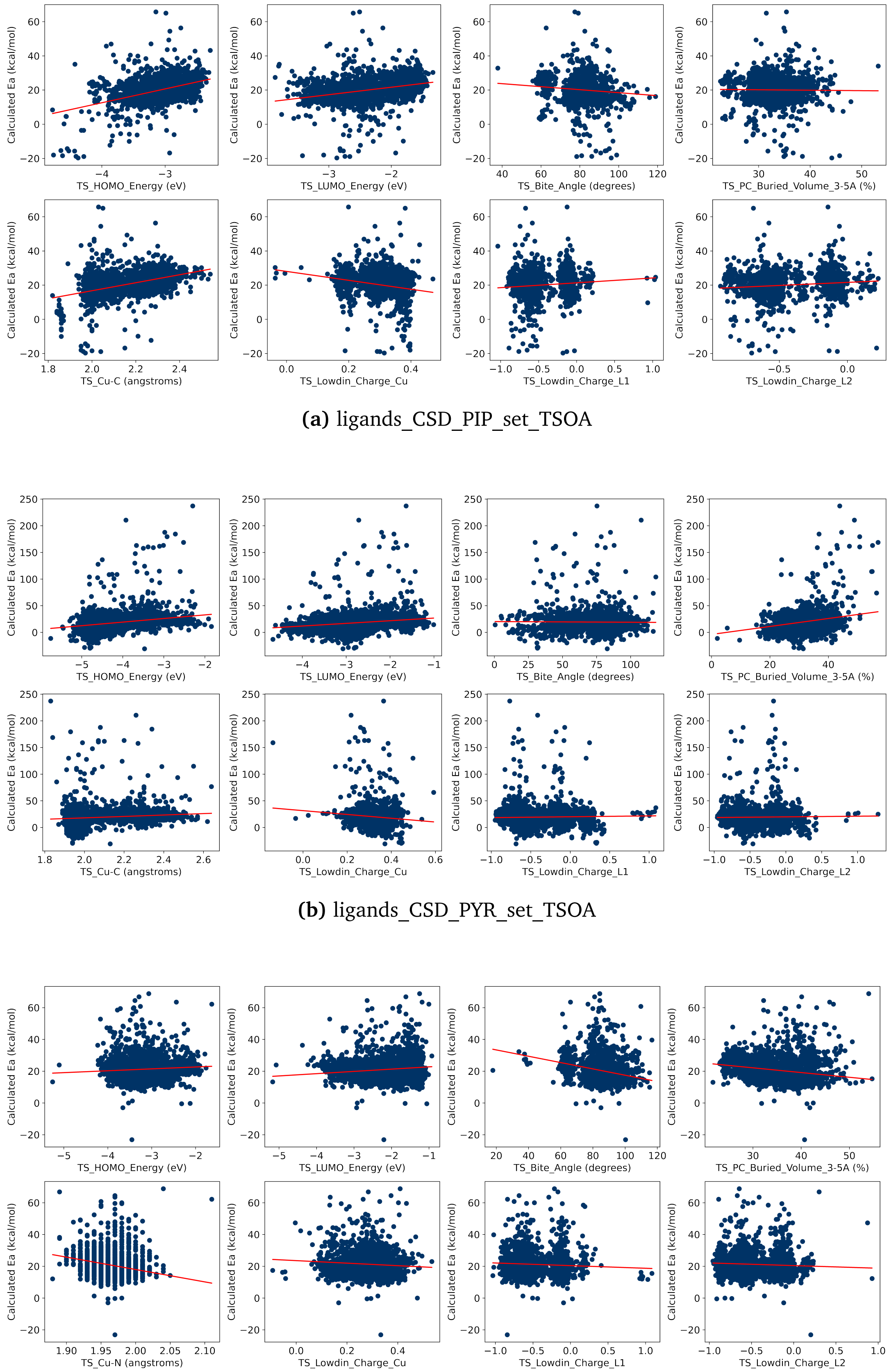
Table 5.1: Full list of descriptors and their source. Entries highlighted in red and blue are for **TSOA** and **TSSig** only respectively. Descriptors 25-31 are extracted for the Cu, nucleophile N, ligand coordinating atom 1 (L1), ligand coordinating atom 2 (L2) and the iodobenzene C and I atoms. For 2-pyrrolidinone the amide C and O atoms are also included for the **TSSig** transition state.

No.	Descriptor	Source	Description
1	Bite_Angle	Python	Ligand bite angle (°)
2	D_Bite_Angle	Python	Change in bite angle from the Cu-I intermediate to the transition state (°)
3	Cone_Angle	Python	Ligand cone angle (°)
4	Sterimol_B1	Python	Smallest distance perpendicular to the Cu-dummy vector to the edge of the ligand (Å)
5	Sterimol_B5	Python	Largest distance perpendicular to the Cu-dummy vector to the edge of the ligand (Å)
6	Sterimol_L	Python	Distance along the Cu-dummy vector to the edge of the ligand (Å)
7	PC_Buried_Volume_3-5Å	Python	Percentage buried volume at a 3.5 Å radius (%)
8	PC_Buried_Volume_5Å	Python	Percentage buried volume at a 5 Å radius (%)
9	PC_Buried_Volume_7Å	Python	Percentage buried volume at a 7 Å radius (%)
10	SASA	Python	Solvent Accessible Surface Area ($Å^2$)
11	HOMO_Energy	ORCA	Energy of the HOMO (eV)
12	LUMO_Energy	ORCA	Energy of the LUMO (eV)
13	Cu-L_x	Python	Bond distance between Cu and ligand atom x (Å)
14	D_Cu-L_x	Python	Change in bond distance between Cu and ligand atom x between the CuLI intermediate and transition state (Å)
15	Cu-I	Python	Cu-I bond distance (Å)
16	Cu-C	Python	Cu-C bond distance (Å)
17	Cu-N	Python	Cu-N bond distance (Å)
18	C-I	Python	C-I bond distance (Å)
19	I-C-Cu	Python	I-C-Cu bond angle (°)
20	N-Cu-I	Python	N-Cu-I bond angle (°)
21	Cu-I-C	Python	Cu-I-C bond angle (°)
22	I-C-N	Python	I-C-N bond angle (°)
23	C-N-Cu	Python	C-N-Cu bond angle (°)
24	C-Cu-I	Python	C-Cu-I bond angle (°)
25	Löwdin_Charge	ORCA	Löwdin atomic charge of the atom
26	Bonded_Valence	ORCA	Number of bonds formed by the atom
27	Atomic_Population	ORCA	Number of electrons localised on the atom
28	Bond_Order	ORCA	Number of bonds between two atoms
29	Orbital_Charge_s	ORCA	Orbital charge of the *s* orbital
30	Orbital_Charge_p and subshells	ORCA	Orbital charge of the *p* orbital and its subshells
31	Orbital_Charge_d and subshells	ORCA	Orbital charge of the *d* orbital and its subshells
32	Img_Freq	ORCA	Magnitude of the imaginary frequency (cm^{-1})

Where an atomic property (e.g. atomic population) is stated the property is extracted for atoms, Cu, ligand coordinating atom 1 (L1), ligand coordinating atom 2 (L2), the nucleophile nitrogen atom and the carbon and iodine atom in the C-I bond in the aryl halide. For the sigma metathesis transition state (**TSSig**) these values were also extracted for the amide carbon and oxygen atom in the 2-pyrrolidinone nucleophile. All descriptors were extracted from the transition state structure except the `D_*` descriptors which are the difference between the transition state and the **CuLI** complex. This resulted in four datasets, ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PIP_set_TSSig***, ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_PYR_set_TSSig*** containing 78, 91, 128 and 130 descriptors and 1683, 3708, 3990 and 5798 ligands respectively.

5.2.2 Correlations Between Descriptors and Activation Energies

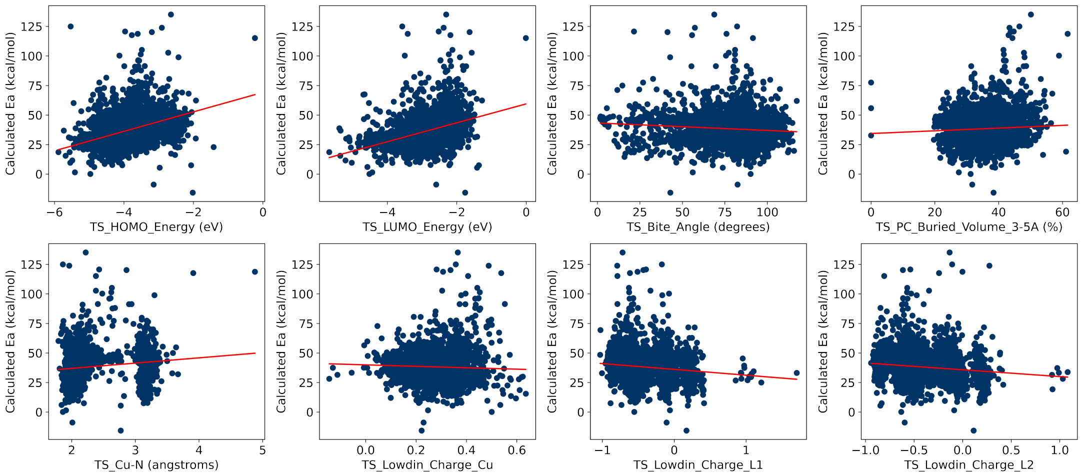
Previous computational studies on Pd-catalysed reactions have shown the dependence of $\Delta G^{\ddagger}$ on the electronic properties of the ligand and its bite angle.[226,227] Thus, an analysis of the properties of the calculated transition states and their relationship to the calculated $\Delta G^{\ddagger}$ in the Ullmann-Goldberg reaction was performed. These properties are: HOMO and LUMO energies of the transition state, cone angle and bite angle of the ligands, % of the buried volume of the ligand around the copper centre, Löwdin charge on Cu and N atoms, Cu−Ph/Cu−N bond length, and atomic population on Cu in the transition state (**Figure 5.4a**). These properties were selected to represent the steric and electronic properties of the catalytic centre in the transition state, which should significantly influence its stability and the calculated $\Delta G^{\ddagger}$. Particular attention was given to the steric descriptors, given the shorter Cu−C/N/O bonds compared to those of palladium. A full breakdown of R^2 values between each descriptor and the activation energy is available in **Appendix 5.A.1**.



(a) ligands_CSD_PIP_set_TSOA

(b) ligands_CSD_PYR_set_TSOA

(c) ligands_CSD_PIP_set_TSSig



(d) ligands_CSD_PYR_set_TSSig

Figure 5.4: Correlation of commonly used steric and electronic descriptors with activation energy for all four datasets.

Surprisingly, no clear relationship is observed with any of the four sets of calculated transition states (***ligands_CSD_Pyr_set_TSOA***, ***ligands_CSD_Pyr_set_TSSig***, ***ligands_CSD_Pip_set_TSOA***, and ***ligands_CSD_Pyr_set_TSSig***). All properties of both the transition states and the starting intermediates have little to no impact on $\Delta G^{\ddagger}$. This highlights the unique nature of Cu(I) d^{10} catalytic centre, which is less sensitive to ligand field geometry and electronic properties of the ligand.

As no straightforward correlation was found between the calculated $\Delta G^{\ddagger}$ and the electronic and steric properties of the transition states of the Ullmann-Goldberg reaction, machine learning was leveraged to probe for more complex relationships between them. While this approach still requires the calculation of the transition states and will not speed up the prediction of $\Delta G^{\ddagger}$, its outcomes may improve our understanding of factors which are important in designing ligands/catalysts for the Ullmann-Goldberg reaction.

5.2.3 Initial Models

5.2.3.1 Data Trimming

All four datasets have an uneven distribution of activation energies with the majority at ~20 kcalmol^{-1}. This will result in poor predictions in the resulting machine learning models due to data bias. To reduce the effect of data bias each dataset was flattened to produce a smoother distribution of activation energies. Activation energies were split into bins of 1 kcal mol^{-1} and randomly samples for a maximum of 100, 200, 250 and 300 entries for the ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig*** datasets respectively. If a bin has less than the maximum number of entries, all entries were kept. **Figure 5.5** shows the distribution of data before and after trimming for the ***ligands_CSD_PIP_set_TSOA*** dataset. The resulting datasets were taken forward to generate the machine-learning models.

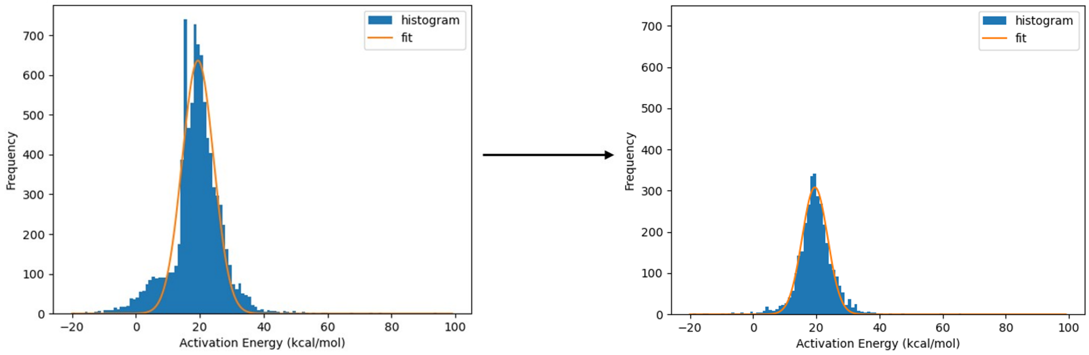


Figure 5.5: Distribution of activation energies before (left) and after (right) trimming for the ***ligands_CSD_PIP_set_TSOA*** dataset.

5.2.3.2 Principal Component Analysis and Linear Regression

Descriptor space was analysed with Principal Component Analysis using the `decomposition.PCA()` method in the scikit-learn Python module and examining the correlation between every descriptor pair using the pandas `corr()` attribute. PCA allows a visual representation of the potential predictive power of 50-60% of the data by plotting the first two principal components. PCA can also be used to find how many descriptors are required to describe a certain amount of variance in the dataset via a scree plot. It should be noted that PCA is not statistically robust, making analyses sensitive to outlier observations.[228]

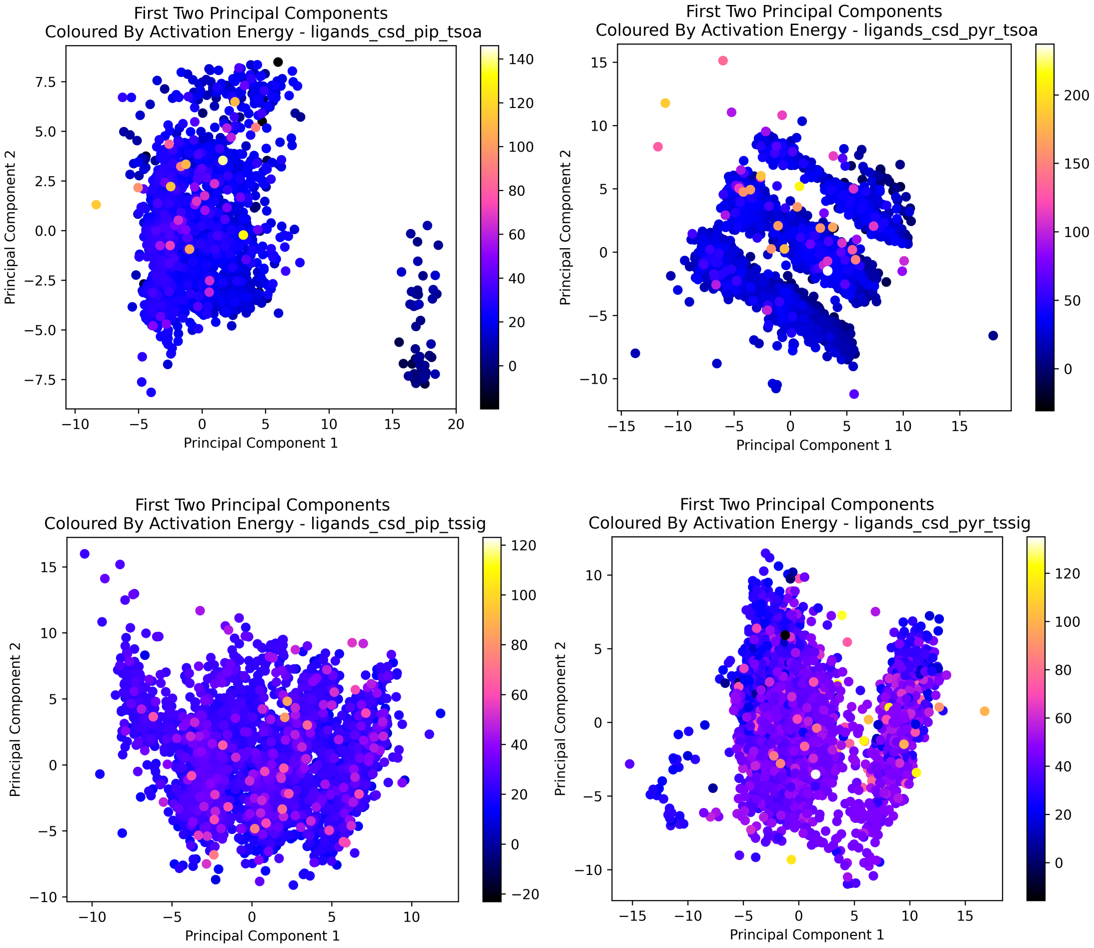
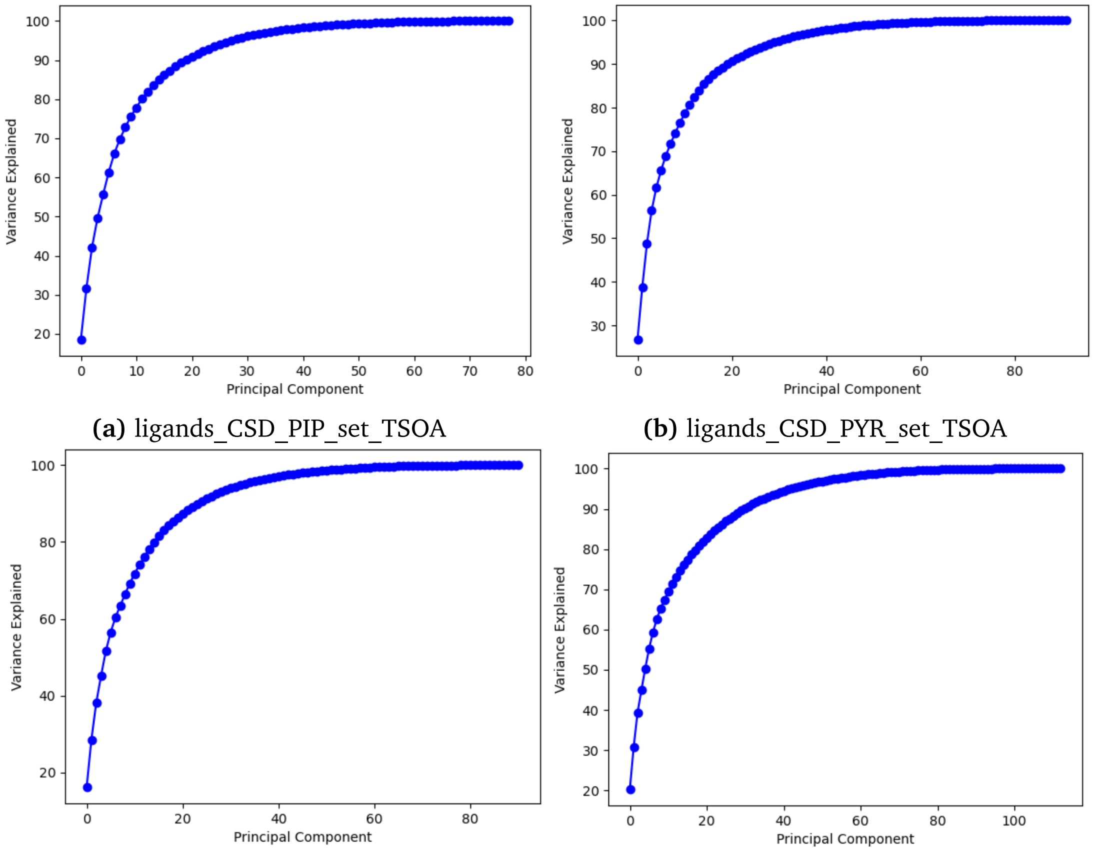


Figure 5.6: Plots of the first two principal components encompassing ∼ 40% of the variance in the data for the four datasets.

The principal component analysis for the first two principal components for all four datasets is shown in **Figure 5.6**. The PCA plots show that there is no obvious relationship between the descriptors and activation energy. This is consistent with the observation that there was no obvious correlation between individual descriptors and activation energy. For ***ligands_CSD_PIP_set_TSOA*** a cluster is observed with a high principal component 1 which has lower activation energies, however, low activation energies are also observed at lower values of principal component 1. This suggests that the correlation between descriptors and activation energy is much more complex and requires a more complex multivariate approach to try and determine the correlation between the structural and electronic properties of the transition states and their respective activation energies.



(a) ligands_CSD_PIP_set_TSOA

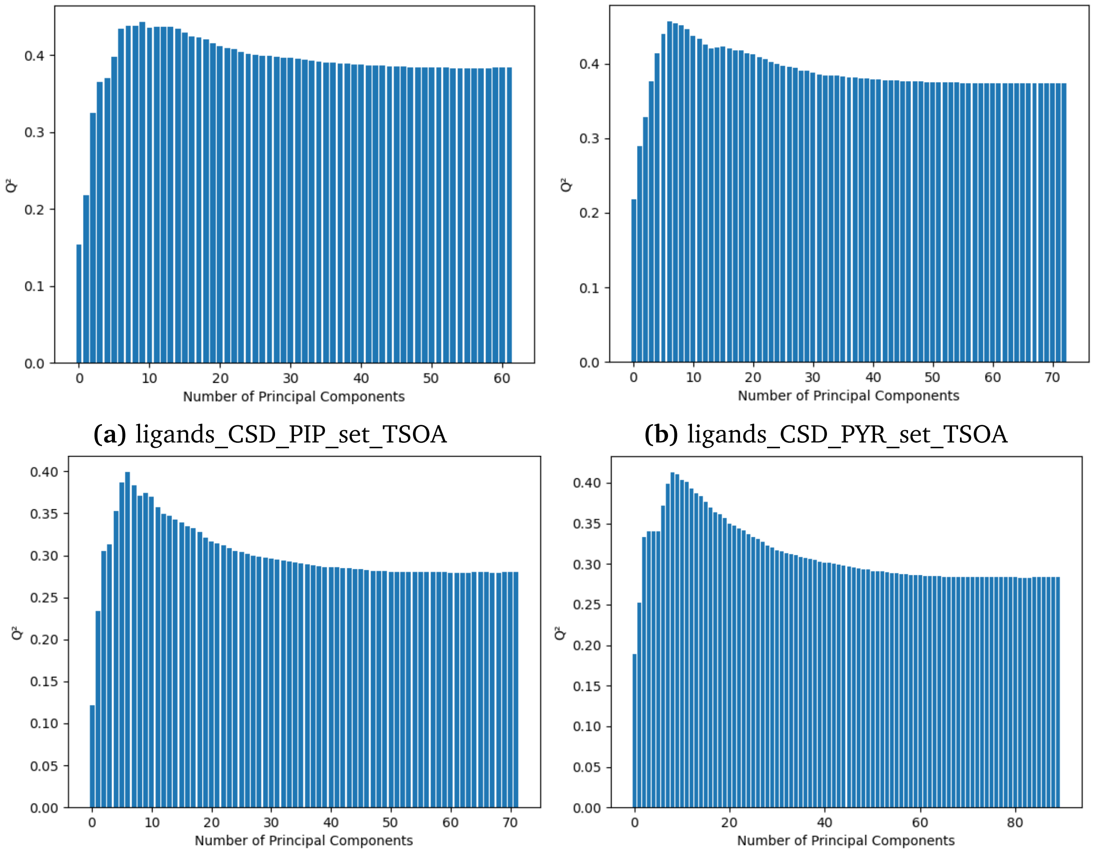
(b) ligands_CSD_PYR_set_TSOA

(c) ligands_CSD_PIP_set_TSSig

(d) ligands_CSD_PYR_set_TSSig

Figure 5.7: Scree plots for the four datasets.

The Scree plot can be used to determine whether there are a large number of correlated or redundant descriptors in the dataset. The scree plots show that after 40 components, most of the variance in the dataset has been described. For ***ligands_CSD_PYR_set_TSSig*** the variance is described in approximately 50 descriptors. This indicates that over half of the descriptors are not necessary.



(a) ligands_CSD_PIP_set_TSOA **(b)** ligands_CSD_PYR_set_TSOA

(c) ligands_CSD_PIP_set_TSSig **(d)** ligands_CSD_PYR_set_TSSig

Figure 5.8: Q^2 plots for the four datasets using 10-fold cross-validation.

The predictability of the dataset can also be assessed using a Q^2 plot using the pcaMethods library in R.[229] Q^2 calculates the ability of 90% of the data to predict the other 10% using 10-fold cross-validation. The higher the Q^2 value, the more consistency is in the data. The amount of variance accounted for is plotted for increasing numbers of components and shown in **Figure 5.8**. The Q^2 plots show that the **TSOA** datasets are slightly more consistent, by 0.05, and may therefore lead to better predictions. For ***ligands_CSD_PIP_-set_TSOA***, ***ligands_CSD_PYR_set_TSOA*** and ***ligands_CSD_PIP_set_TSSig*** only 8 components are needed to make an effective prediction. For ***ligands_CSD_PYR_set_TSSig*** 14 components are required. This also demonstrates that there are a lot of redundant descriptors in the datasets. Interestingly for all datasets the Q^2 value decreases after the peak in the optimum number of components. This suggests that there are only a small number of descriptors that are responsible for the prediction and multiple descriptors that are negatively affecting the predictions and need to be identified and removed. In order to determine which descriptors can be removed, preliminary machine learning models were generated and the importance of the descriptors was determined.

5.2.3.3 Choice of Machine Learning Methods

Eight machine learning methods were chosen with varying complexity, ranging from simple Multiple Linear Regression to more advanced Neural Networks. The eight initial methods chosen were Multiple Linear Regression (MLP), Gaussian Process (GP), Artificial Neural Network (ANN), Partial Least Squares (PLS), Support Vector Machine (SVM), Random Forest (RF), ExtraTrees (ET) and Bagging (Bag). Artificial Neural Networks use one hidden layer due to the size of the dataset and to reduce computational time. Parameters were optimised for all models except MLR (see **Section 5.4.2**).

5.2.3.4 Metrics & Validation

The coefficient of determination (R^2) and root mean squared error (RMSE) were used to assess the models. Two new metrics were created for evaluation to compare predicted values against the computationally calculated activation energies: % of predictions within ± 4 kcal mol^{-1} and within ± 2 kcal mol^{-1} (% within 4.0 and % within 2.0). The former reflects the maximum accuracy of the model as 4 kcal mol^{-1} is the average error of the energies generated from the computational workflow against DLPNO-CCSD(T)/def2-TZVPP energies (see **Section 3.2.7**). The latter is the percentage within half the error of the underlying data.

Models were assessed with explicit training and test sets. The training and test sets were created by grouping the activation energies into bins of 1 kcal mol^{-1} and randomly splitting each bin in an 80:20 split to form the training and test sets respectively.

5.2.3.5 Descriptor Correlations

Analysis of the correlation between descriptors is a key phase in pre-processing machine learning datasets. If two descriptors are highly correlated it is indicative of a causal relationship. Therefore the two descriptors describe the same property and one of the two descriptors can be removed. It is important to note however that correlation does not equal causation and the nature of each descriptor must also be considered. A high correlation between bond order and bond length is highly likely to be describing the same property of the complex, whereas a high correlation between an electronic descriptor and a steric descriptor is not likely describing the same property. Trimming highly correlated descriptors decreases the noise within the resultant models as well as decreases the training time due to a lower number of features to fit.

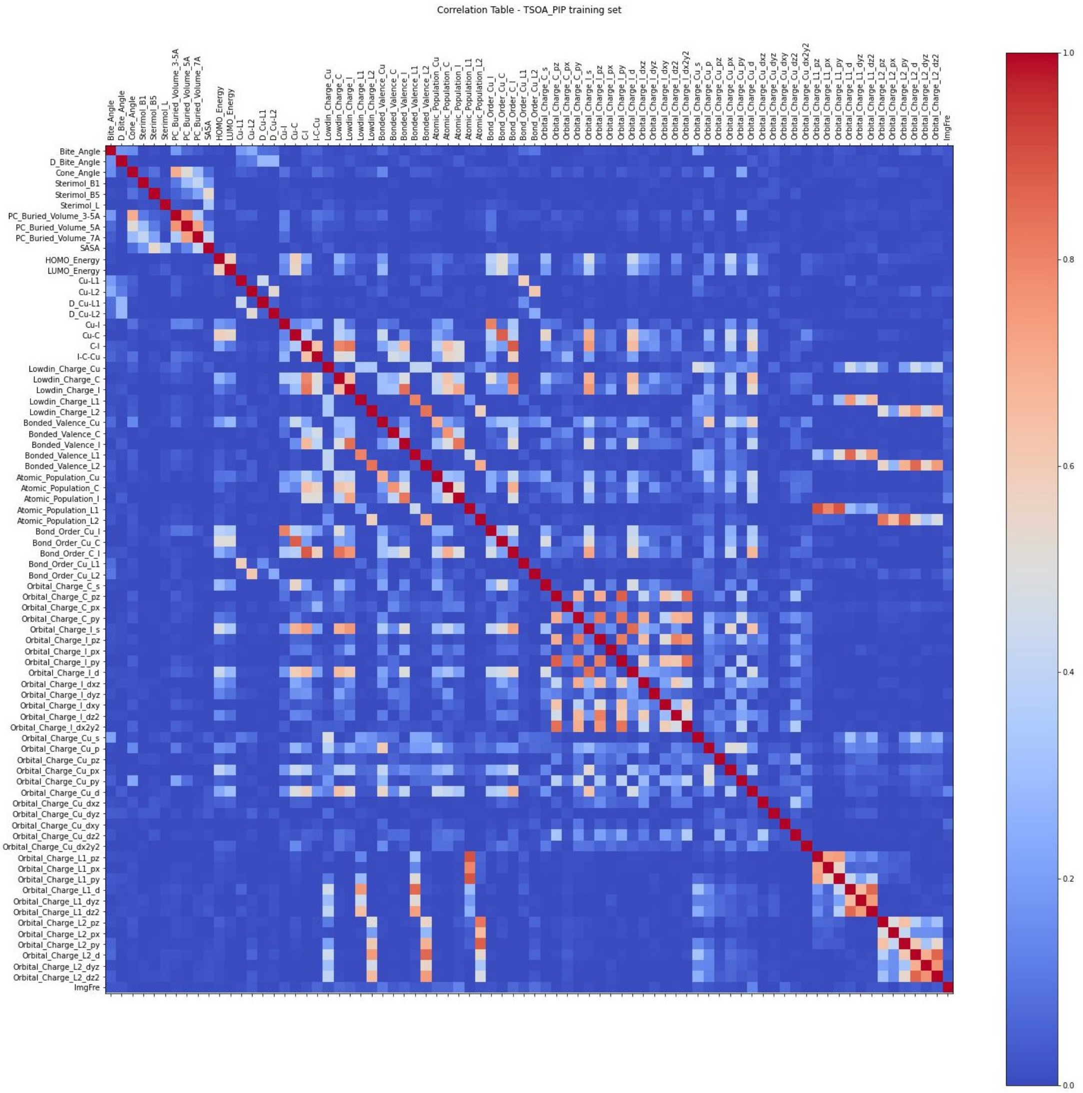


Figure 5.9: Correlation plot between all the descriptors in the ***ligands_CSD_PIP_set_-TSOA*** dataset after trimming.

For descriptors with a correlation >0.9 and describing similar fundamental properties, one of the two descriptors was removed from the dataset. An example descriptor correlation plot using R^2 for ***ligands_CSD_PIP_set_TSOA*** is shown in **Figure 5.9**. The analysis for the other three datasets is almost identical. Common descriptors that are highly correlated across all four datasets are bond lengths and bond angles, bond length and bond order, Löwdin charge an*d* orbital charge, atomic population an*d* orbital charge an*d* orbital charges between orbitals and subshells (e.g. d and d_{xy}). Descriptors highlighted in red ($R^2 > 0.9$, **Table 5.2**) were removed from the dataset.

Table 5.2: Correlation of descriptors for the ***ligands_CSD_PIP_set_TSOA*** dataset where $R^2 > 0.9$. Descriptors highlighted in red were removed.

Descriptor 1	Descriptor 2	R^2
Cu-C (Å)	Cu-I-C (°)	0.92
C-I (Å)	C-Cu-I (°)	0.97
Löwdin Charge (C)	Orbital Charge C(p)	0.98
Löwdin Charge (I)	Orbital Charge I(p)	0.98
Atomic Population (L1)	Orbital Charge L1(s)	0.98
Atomic Population (L1)	Orbital Charge L1(p)	0.98
Atomic Population (L2)	Orbital Charge L2(s)	0.97
Atomic Population (L2)	Orbital Charge L2(p)	0.97
Orbital Charge L1(s)	Orbital Charge L1(p)	0.97
Orbital Charge L1(p)	Orbital Charge L1(p$_z$)	0.91
Orbital Charge L1(d)	Orbital Charge L1(d$_{xz}$)	0.94
Orbital Charge L1(d)	Orbital Charge L1(d$_{xy}$)	0.90
Orbital Charge L1(d$_{xy}$)	Orbital Charge L1(d$_{x^2-y^2}$)	0.91
Orbital Charge L2(s)	Orbital Charge L2(p)	0.96
Orbital Charge L2(d)	Orbital Charge L2(d$_{xz}$)	0.94
Orbital Charge L2(d)	Orbital Charge L2(d$_{xy}$)	0.91
Orbital Charge L2(d)	Orbital Charge L2(d$_{x^2-y^2}$)	0.91
Orbital Charge L2(d$_{xy}$)	Orbital Charge L2(d$_{x^2-y^2}$)	0.92

Removal of highly correlated descriptors resulted in a set of 75 descriptors for ***ligands_CSD_PIP_set_TSOA***, 90 descriptors for ***ligands_CSD_PIP_set_TSSig***, 111 descriptors for ***ligands_CSD_PYR_set_TSOA*** and 121 descriptors for ***ligands_CSD_Pyr_set_TSSig***. These datasets were used to generate the initial machine-learning models.

5.2.3.6 Initial Predictions

Models were assessed using explicit training and test sets. The test sets were created by randomly removing data (evenly across the activation energy distribution) from the datasets. The remaining data comprises the training sets. The distribution of data for the four datasets is shown in **Table 5.3**. The ***ligands_CSD_PIP_set_TSOA*** dataset has a significantly lower number of data points due to the low success rate of the **TSOA** transition state with the piperidine nucleophile.

Table 5.3: Size of the dataset (N), training set (T) and test set (S) used for the initial predictions.

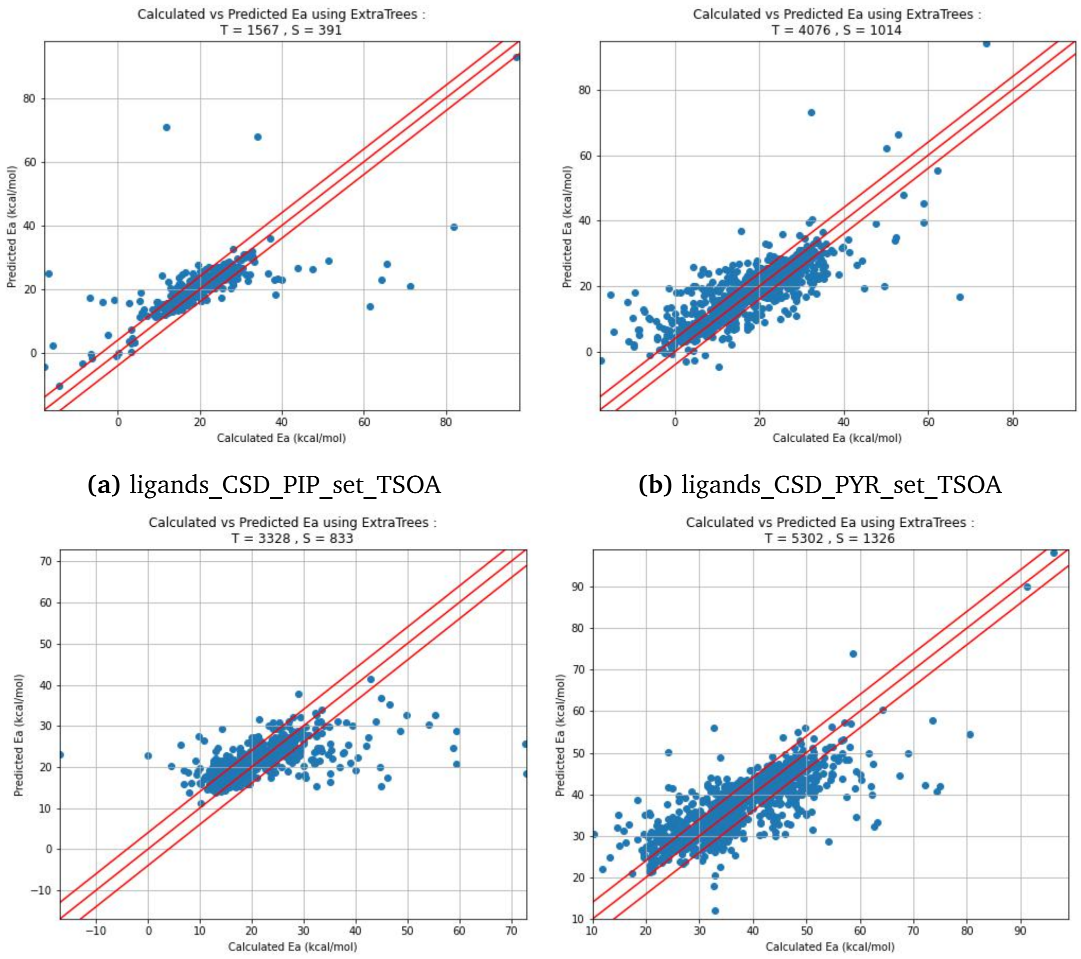
Dataset	N	T	S
ligands_CSD_PIP_set_TSOA	1963	1570	393
ligands_CSD_PYR_set_TSOA	5095	4076	1019
ligands_CSD_PIP_set_TSSig	4161	3328	833
ligands_CSD_PYR_set_TSSig	6628	5302	1326

For all datasets, the ExtraTrees method performed the best (**Figure 5.4**). Tree-based methods (RF and Bagging) as well as SVM gave comparable results with linear methods (MLR and PLS) as well as neural networks (ANN) and Gaussian Process (GP) gave significantly poorer results. Initial models for the piperidine nucleophile (***ligands_CSD_PIP_set_TSOA*** and ***ligands_CSD_PIP_set_TSSig***) have poor R^2 values <0.5 but have reasonably good values of % within 4.0 (75.5-77.9). The opposite is true for the 2-pyrrolidinone nucleophile, R^2 values are acceptable at ∼0.65, but % within 4.0 is lower at 66.1-68.5. RMSE across all datasets, however, is quite poor with values much higher than the calculated error in the dataset (3.9 $\text{kcal}\,\text{mol}^{-1}$). It is also observed that increasing the number of data points generally improves the RMSE. ***ligands_CSD_PIP_set_TSOA*** with a significantly lower number of data points (∼2000) has much poorer RMSE at 7.90 compared to the other three datasets (5.52-6.32). Full results with all machine learning methods for all datasets are shown in **Appendix 5.A.4**.

Table 5.4: Metrics of the best machine learning methods for each dataset using explicit training and test sets. The best method is defined as the method which produced the majority of best metrics. At least three out of four metrics were the best for the methods displayed.

Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA	ET	0.49	7.90	75.5	54.0
ligands_CSD_PYR_set_TSOA	ET	0.65	6.32	66.1	45.8
ligands_CSD_PIP_set_TSSig	ET	0.39	5.93	77.9	59.9
ligands_CSD_PYR_set_TSSig	ET	0.63	5.52	68.5	46.1

The predicted vs calculated activation energies were plotted to visualise the models. The data is shown in **Figure 5.10** for ExtraTrees only.



(a) ligands_CSD_PIP_set_TSOA

(b) ligands_CSD_PYR_set_TSOA

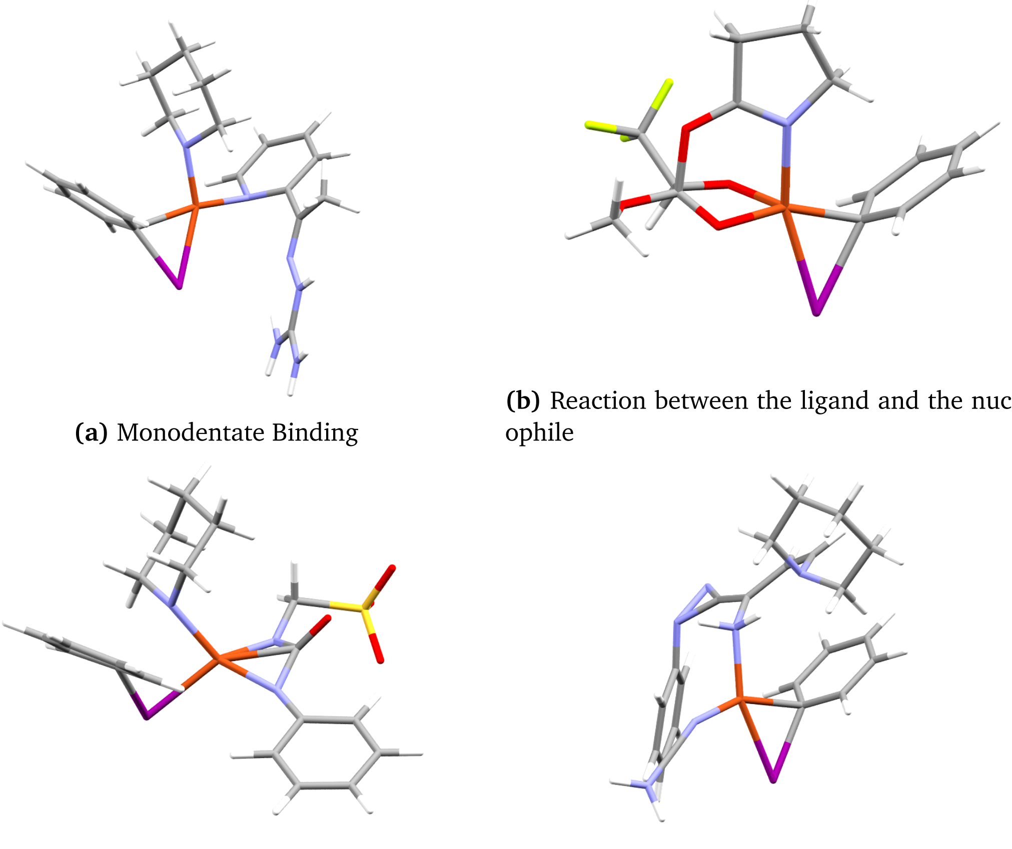
(c) ligands_CSD_PIP_set_TSSig

(d) ligands_CSD_PYR_set_TSSig

Figure 5.10: Calculated activation energies vs predicted activation energies for all four datasets using the ExtraTrees model.

Visualisation of the models shows that around the mean (~20 kcal mol^{-1}) predictions are generally very good, with the majority of data points within ±4.0 kcal mol^{-1}. However, predictions are much poorer at low and high activation energies. Predictions at <10 kcal mol^{-1} and >35 kcal mol^{-1} (>55 kcal mol^{-1} for ***ligands_CSD_PYR_set_TSSig***) are generally very poor. Predictions at activation energies >40 kcal mol^{-1} have errors >8 kcal mol^{-1}. This is likely due to the lack of training data in these ranges (see **Section 5.2.3.1** and **Appendix 5.A.2**).

The outliers over all models were examined and were defined as a prediction with an error >6 kcal mol^{-1} (1.5× RMSE). Structures at low activation energies often possess less common chemistry such as a hydrogen bond between the ligand and the nucleophile N or O atoms. Such chemistries are present in low quantities in the training sets and are not accounted for in the descriptor set. The poor description and lack of suitable training data containing the same chemistry is the most likely cause for the poor predictions. The poor predictions at higher activation energies were initially thought to be due to the lack of data in this range of activation energies, however, upon inspection of the structures, this was not the case. In most cases, the structures of the transition states for these outliers were incorrect and not picked up during the structure screening process during calculation. High activation energy outlier structures either had incorrect coordination, binding in a monodentate manner instead of a bidentate manner, incorrect structure of the ligand or unusual binding mode such as a 4-membered ring which is under-represented in the training data.



(a) Monodentate Binding

(b) Reaction between the ligand and the nucleophile

(c) Four-membered ring coordination

(d) Poor complex structure

Figure 5.11: Examples of common structural issues in outliers.

All models accurately predict incorrect structures as outliers and can be used as an effective way of identifying incorrect structures within the datasets without the need to manually examine each structure individually. This demonstrates that machine learning can be used to identify incorrect structures that are not picked up during the high-throughput calculation workflow. For large datasets containing >1000 structures, a manual examination is extremely time-consuming, therefore, being able to automatically identify incorrect structures is extremely valuable. All outliers at activation energies >40 were manually inspected and incorrect structures were removed. To remove structures with one ligand coordinating atom dissociated all entries were removed with a $\mathrm{Cu{-}L_1}$ or $\mathrm{Cu{-}L_2}$ bond order of 0. The resulting data sets were used to optimise the models.

5.2.4 Parameter Optimization

All parameters were optimised, using the Optuna Python module, to maximise the Coefficient of Determination (R^2).[230] Only R^2 is maximised, therefore other metrics may be worse with an optimised parameter set. It is also important to note that the parameters are only optimised on the training set and therefore, may perform worse on the test set. In general, optimised parameters should give better predictions ensuring that the range of values of each parameter explored is within a suitable range. For all datasets and models, R^2 increased when optimised parameters were used. The optimised parameters were retained for use in evaluating each model on an explicit training and test set. 10-fold cross-validation was used to ensure all data points were used at least once during optimisation.

5.2.4.1 Gaussian Process

The kernel (also called the 'covariance function') is a crucial part of a Gaussian Process model. The kernel encodes the assumptions on the function being learned by defining the similarity of two data points combined with the assumption that similar data points should have similar target values. Only the stationary kernels Matern, Radial Basis Function (RBF) and RationalQuadratic were tested. The RationalQuadratic kernel was the best-performing kernel for all datasets.

5.2.4.2 Artificial Neural Network

Due to the small size of the datasets (1600-5800 ligands), all artificial neural networks were limited to a single layer in this section. While the number of data points is within an acceptable amount for simple Deep Neural Networks the accuracy of neural network-based models was assessed first with a single-layer artificial neural network. The number of nodes in the single hidden layer `n_estimators` was optimised to maximize the value of R^2 at the lowest computational cost (increasing the number of nodes also increases the time taken to train the model). At a low number of nodes, `n_estimators` <100 the models improve as the number of nodes increases. Several networks in this range also fail to optimise entirely. After 400 nodes the metrics become worse due to over-fitting as seen in **Figure 5.12**. After optimization across all four datasets, the value of `n_estimators` was set to 400, 800, 800 and 700 for the ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig*** datasets respectively. Full results with all metrics across the four datasets are shown in **Appendix 5.A**.

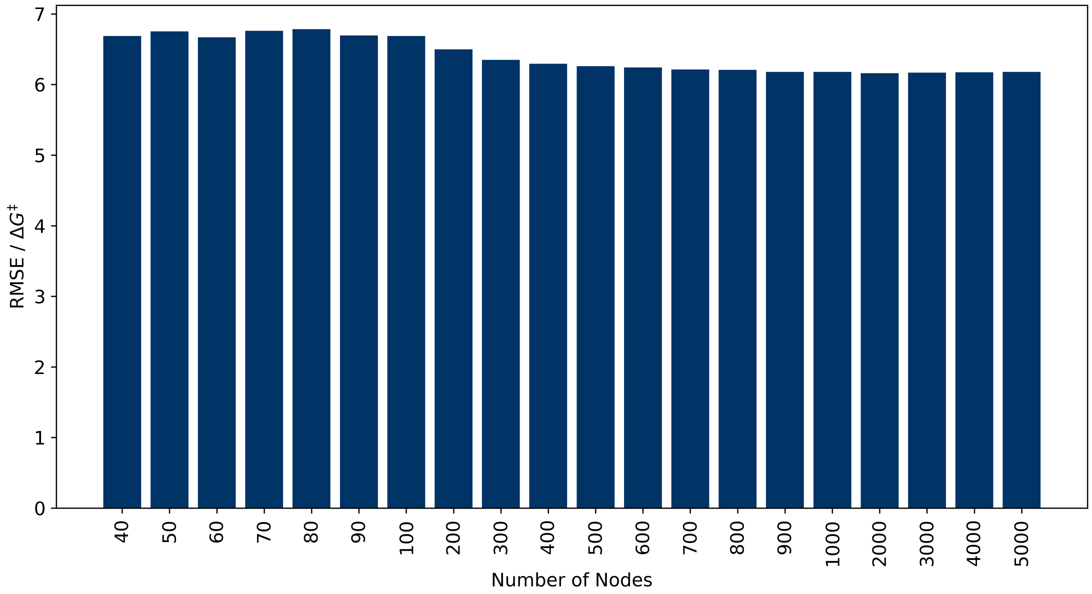


Figure 5.12: The effect of changing the number of nodes in a single hidden layer on the RMSE during training on the performance of the ***ligands_CSD_PIP_set_TSSig*** test set.

5.2.4.3 Support Vector Machine

SVM has several parameters which must be carefully optimised to give the best fit. The 'kernel' (the function used to transform non-linearity to linearity); 'c' (the penalty parameter on the error term); 'epsilon' (which specifies the limits whereby no penalty is associated in the training loss function) and 'gamma' (the kernel coefficient). The optimised SVM parameters for each dataset are shown in **Table 5.5**.

Table 5.5: Optimised values of c, epsilon and gamma for the four datasets.

Dataset	c	epsilon	gamma
ligands_CSD_PIP_set_TSOA	1328	1e-10	0.0009
ligands_CSD_PYR_set_TSOA	78	2.64	0.008
ligands_CSD_PIP_set_TSSig	175	0.03	0.004
ligands_CSD_PYR_set_TSSig	566	0.0003	0.0006

5.2.4.4 Partial Least Squares

The partial least square method implemented in scikit-learn is 'ready to go' in its default implementation. PLS reduces the number of dimensions of the datasets, which is good for models using many redundant descriptors. The only parameter that needs to be optimised is the number of components (`n_components`), which is the number of predictors to keep after the reduction of the number of descriptors. `n_components` equals 13, 12, 37 and 99 for ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_-PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig*** datasets respectively.

5.2.4.5 Tree Methods

Tree methods are often 'ready to go' in their default implementation. The only parameter which needs to be optimised is the number of trees `n_estimators`. In most cases increasing the number of trees leads to better predictions.[231] To prevent overfitting and to reduce the computational time, the number of trees was varied between 1 and 5000. This is shown in **Figure 5.13** for ExtraTrees optimization on the ***ligands_CSD_PIP_set_TSSig*** dataset, where the change in RMSE after 100 trees is negligible. Due to the similarity between Random Forest and the ExtraTrees and Bagging methods, the same trend was seen for all models across the four datasets. For the initial models 200, 600, 400 and 300 trees were the standard values for the ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PYR_-set_TSOA***, ***ligands_CSD_PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig*** datasets respectively for all tree models.

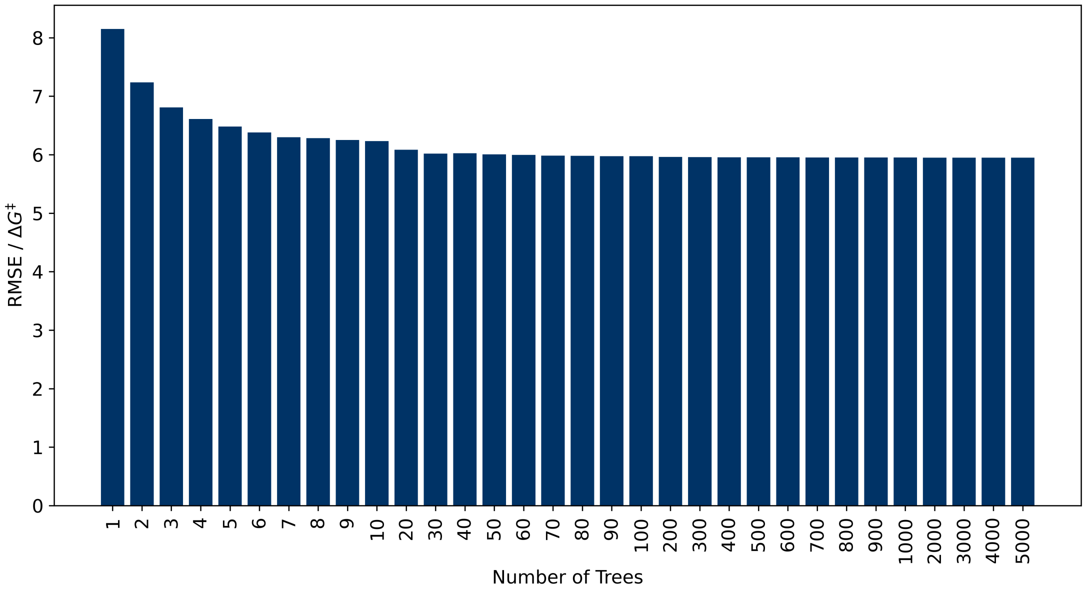


Figure 5.13: The effect of changing the number of trees on the RMSE during training on the performance of the ***ligands_CSD_PIP_set_TSSig*** test set.

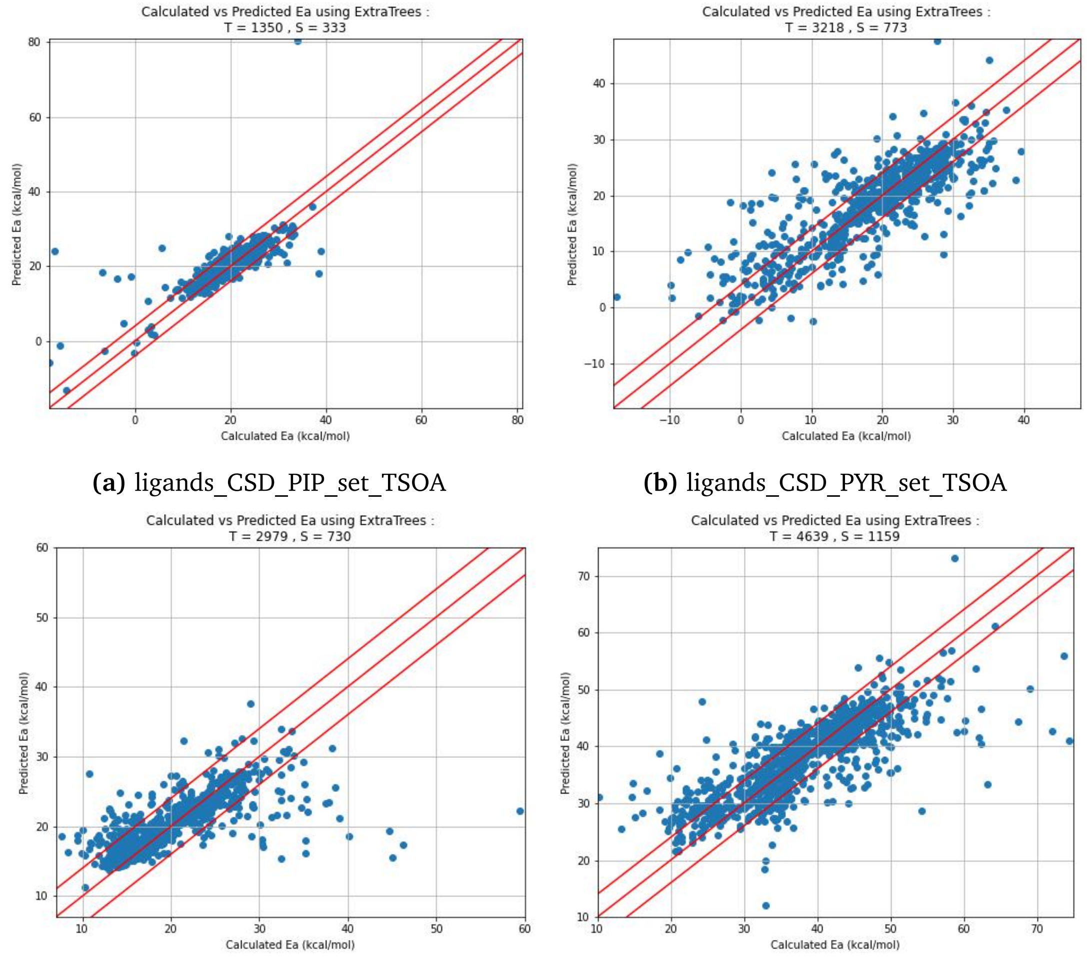
5.2.4.6 Models with Optimised Hyperparameters

Optimisation of hyperparameters as well as the removal of outliers shows a large improvement in model performance. ExtraTrees remained the best-performing model for all four datasets. All metrics improve across all four datasets. R^2 values improve by 0.1 for the **PIP** datasets with a slight improvement for the **PYR** datasets (0.02-0.03). RMSE values also show a large improvement (0.5-2.65), especially for ***ligands_CSD_PIP_set_TSOA*** with an improvement of 2.65 kcal mol^{-1}. % within 4.0 and 2.0 also showed improvements of approximately 5% for all datasets. The removal of outliers with incorrect structures is likely the most important factor in the substantial improvement in the models. Removal of outliers with errors >8 kcal mol^{-1} is expected to improve all metrics by a large margin. The effect of the optimisation of hyperparameters was found to be 0.01-0.05 kcal mol^{-1} across all models, when compared to the same data without the outliers removed.

Table 5.6: Metrics of the ExtraTrees models with optimised hyperparameters for all four datasets with outliers removed.

Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA	ET	0.59	5.25	80.2	55.3
ligands_CSD_PYR_set_TSOA	ET	0.68	5.13	71.2	50.1
ligands_CSD_PIP_set_TSSig	ET	0.48	4.33	82.6	62.9
ligands_CSD_PYR_set_TSSig	ET	0.65	5.02	71.6	47.7

Visualisation of the models (**Figure 5.14**) shows that distribution of the predicted activation energies is much improved over the initial models with the large majority of data lying in the ±4 kcal mol^{-1} boundaries. There are fewer extreme outliers across all datasets. All high activation energy outliers were removed for both **TSOA** datasets. This demonstrates the ability of the models to identify incorrect structures. For all models, activation energies at <10 kcal mol^{-1} were worse across all datasets and high activation energy outliers for the **TSSig** datasets. Predictions at activation energies >40/55 kcal mol^{-1} still have large errors, but correct structures, and can now be attributed to the lack of suitable training data in these activation energy ranges.



(a) ligands_CSD_PIP_set_TSOA

(b) ligands_CSD_PYR_set_TSOA

(c) ligands_CSD_PIP_set_TSSig

(d) ligands_CSD_PYR_set_TSSig

Figure 5.14: Calculated activation energies vs predicted activation energies for all four datasets using the ExtraTrees model with optimised hyperparameters.

To further improve the models the importance of each descriptor was calculated in order to determine the number of redundant descriptors, as well as any descriptors negatively impacting predictions.

5.2.5 Feature Importance

Having optimised the parameters, further model improvements can be achieved by optimising the number of descriptors used in the models. From the PCA analysis (**Section 5.2.3.2**) it was observed that a large number of descriptors were not required to give describe most of the variance in the dataset. It was also observed that several descriptors may be negatively affecting the predictions.

Tree-based models provide a measure of descriptor importance based on the mean decrease in impunity (MDI). Impunity is quantified by the splitting criteria of the decision trees. This method, however, can give high importance to features that may not be predictive for unseen data if the model is overfitting. Impunity-based importances are also strongly biased, and favour numerical features over binary or categorical features. Permutation importance is an alternative importance method that avoids this issue as it can be calculated on unseen data and does not exhibit descriptor-type bias. To assess the importance of each descriptor on the models scikit-learn's `permutation_importance()` was used for the ExtraTrees method.

Permutation importance is a model inspection technique that can be used for non-linear estimators (e.g. trees, neural networks) with tabular data. The permutation descriptor importance is defined as the decrease in the model score when a single descriptor is randomly shuffled. Shuffling the descriptor breaks the relationship between the descriptor and the output variable, thus the model score is indicative of how much the model depends on the feature. Permutation importance is not reflective of the intrinsic predictive power of the descriptor by itself but how important it is for a particular model.

Permutation importance can be calculated on either the training set or the test set. Descriptors that are important on the training set but not on the test set may cause the model to overfit. Whereas using a test set makes it possible to highlight which descriptors contribute the most to the generalisation power of the model. Permutation importance was calculated for the test sets of all four datasets for the ExtraTrees models. Descriptor importance was scored based on the coefficient of determination (R^2), negative mean absolute percentage error and negative mean squared error. Importance's are reported as the mean decrease in the scoring function ± the standard deviation over 50 repeats. The random state was kept constant to ensure the same data was used for each scoring method and for comparison across models. As highly correlated descriptors were removed previously, issues with highly correlated features providing the same information during shuffling, reducing the importance of both features, should not be present.

Permutation importance was calculated for all descriptors to determine their importance in the prediction. Descriptors with a $mean - 2 * std > 0$ were retained. To generate a consistent descriptor set the retained descriptors were combined for the same transition state. For the **PYR** datasets additional descriptors were added to describe the amide. This resulted in trimmed descriptor sets containing 20 for ***ligands_CSD_PIP_set_TSOA***, 24 for ***ligands_CSD_PYR_set_TSOA***, 36 for ***ligands_CSD_PIP_set_TSSig*** and 41 for ***ligands_CSD_PYR_set_TSSig*** descriptors.

5.2.6 Improved Models with Trimmed Descriptors

Models were rebuilt using the trimmed descriptor sets for all eight machine-learning models. Hyperparameters were optimised using the Optuna python package to fine-tune the hyperparameters.[230] The best performing models for each dataset are shown in **Figure 5.7**.

Table 5.7: Metrics of the best-performing models with optimised hyperparameters for all four datasets with the optimised descriptor sets.

Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA	ET	0.66	4.81	79.6	58.6
ligands_CSD_PYR_set_TSOA	ET	0.71	4.86	71.3	51.2
ligands_CSD_PIP_set_TSSig	ET	0.48	4.33	81.5	62.6
ligands_CSD_PYR_set_TSSig	ET	0.66	4.95	70.6	47.5

Metrics for ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig*** are mostly unchanged with slight improvements in R^2 (0.01-0.02) and RMSE (0-0.27). Therefore, a large number of descriptors in these datasets were redundant and not contributing to the predicting power of the models. For ***ligands_CSD_PIP_set_TSOA***, metrics improve by a significant margin with R^2 and RMSE improving from 0.59 and 5.25 to 0.66 and 4.81 respectively. This suggests that one or more descriptors were negatively affecting the predictions in this dataset. Removal of descriptors across all four datasets resulted in smaller datasets and improved metrics. These smaller descriptor sets mean that fewer descriptors need to be calculated along with reduced time to build the models for the same predictive power.

5.2.7 Important Ligand Properties

Permutation importance was used on the optimised ExtraTrees models with the trimmed descriptor sets to identify the most important ligand properties driving the prediction of the activation energy. The top five descriptors for each transition state and nucleophile are shown in **Table 5.8**.

Table 5.8: Top five most important descriptors, by the reduction in R^2 for the ExtraTrees model for each dataset.

	ligands_CSD_PIP_set_TSOA		***ligands_CSD_PYR_set_TSOA***	
No.	Descriptor	Reduction in R^2	Descriptor	Reduction in R^2
1	HOMO_Energy	0.240 ± 0.022	HOMO_Energy	1.032 ± 0.068
2	Bonded_Valence_I	0.235 ± 0.039	Orbital_Charge_I_d	0.232 ± 0.020
3	Orbital_Charge_I_d	0.187 ± 0.018	Orbital_Charge_Cu_s	0.137 ± 0.010
4	Löwdin_Charge_C	0.135 ± 0.010	Atomic_Population_L2	0.123 ± 0.011
5	Orbital_Charge_Cu_d	0.131 ± 0.025	Orbital_Charge_Cu_d	0.098 ± 0.012
	ligands_CSD_PIP_set_TSSig		***ligands_CSD_PYR_set_TSSig***	
No.	Descriptor	Reduction in R^2	Descriptor	Reduction in R^2
1	Bite_Angle	0.070 ± 0.009	HOMO_Energy	0.220 ± 0.016
2	Atomic_Population_Cu	0.058 ± 0.012	Atomic_Population_Cu	0.074 ± 0.007
3	LUMO_Energy	0.047 ± 0.006	Amide_C-O	0.038 ± 0.005
4	C-N-Cu	0.037 ± 0.009	LUMO_Energy	0.021 ± 0.004
5	Orbital_Charge_I_s	0.035 ± 0.011	Bond_Order_Cu-N	0.020 ± 0.003

For every dataset except ***ligands_CSD_PIP_set_TSSig***, the HOMO energy is the most important property. Bite angle is the most important property for the ***ligands_CSD_PIP_set_-TSSig*** dataset, however, the reduction in R^2 is small (0.070). The difference between the most important descriptors is small, therefore, for ***ligands_CSD_PIP_set_TSSig*** there is a more complex mixture of properties dictating the activation energy for the sigma metathesis transition state with an amine nucleophile. Important descriptors for the oxidation addition transition state (**TSOA**) are the HOMO energy, orbital charge on the copper d orbital and the orbital charge on the iodine d orbital. This suggests that the ligand's ability to influence the electron-withdrawing or donating ability of the copper centre, through the copper d orbital, to the iodine atom during oxidative addition is an important factor in the activation energy of the reaction. Comparing nucleophiles, piperidine has higher importance for the bonded valence of the iodine atom and the charge on the aryl carbon atom, whereas, for 2-pyrrolidinone the charge on the copper *s* orbital and atomic population on the ligand coordinating atoms are important properties. The presence of the amide in the nucleophile may prefer ligands that are able to modulate the electron density of the copper *s* orbital through the ligand coordinating atoms. The difference in electron density at the copper centre may promote the oxidation addition of the aryl halide when an amide is bound to the copper. For piperidine, no such importance is seen in favour of descriptors localised on the aryl halide. The importance of the bonded valence of iodine and the charge on the aryl carbon atom suggests that for piperidine the progress of the C−I bond break at the transition state is key in predicting the activation energy. This suggests that ligands that influence the position of the transition state in the C−I bond-breaking process are important for amine nucleophiles. The difference in bonding to copper between an amine and amide may be the cause for this deviation in ligand requirements in the oxidative addition transition state.

For the sigma metathesis transition state, the atomic population at the copper centre and LUMO energy are present in both nucleophiles, suggesting that the ligand's ability to modulate the electron density at the copper centre is important in this transition state. The amide C−O bond length and Cu−N bond order are also important properties of the amide nucleophile. This implies that the ability of the copper to bond to and weaken the amide bond in the nucleophile is an important factor. A ligand that modulates the electron density at the copper centre, weakening the amide bond may prove to be a more effective ligand. For piperidine, no such trend is observed. Like in the oxidative addition the important descriptors suggest that the activation energy for amine nucleophiles is dictated by the addition of the aryl halide rather than the activation of the nucleophile.

While these feature importance's do not provide a direct trend between property and activation energy, they do provide an insight into the processes that may be taking place. This provides a starting hypothesis that can be further explored either computationally or experimentally.

5.2.8 Transition State Independent Descriptors

The ability to predict the activation energy of a ligand directly from the active catalytic state, without the need to calculate the transition state would be ideal. Calculation of the transition state is the most time-consuming process in the computational workflow. Being able to remove this step and predict the activation energy directly from the active catalytic state, will both reduce the total amount of computing resources required to screen ligands and reduce the total time required to generate a prediction. While complete removal of transition state calculations is not possible, due to them being required to calculate the activation energies needed to train the models, it can reduce the number of total calculations required by a significant amount. Especially if a model is built using a set amount of training data and then used purely as a predictive tool.

To build these models a set of transition state independent descriptors was created using descriptors taken from the active catalytic state (**CuLNu**), which is a stable intermediate. The same descriptor set was used as a base and the descriptors relating to the transition state (e.g. aryl halide) were removed. The remaining descriptors were calculated from the **CuLNu** complex. Descriptors for individual atoms were extracted for the Cu, nucleophile N, ligand coordinating atom 1 (L1) and ligand coordinating atom 2 (L2) atoms. For 2-pyrrolidinone descriptors were also extracted for the amide C and O atoms. A full list of transition state independent descriptors is shown in **Table 5.9**.

Table 5.9: Full list of descriptors for the TS independent descriptor sets and their source. Descriptors are calculated from the CuLNu active catalytic state. Descriptors 14-20 were extracted for the Cu, nucleophile N, ligand coordinating atom 1 (L1) and ligand coordinating atom 2 (L2) atoms.

No.	Descriptor	Source	Description
1	Bite_Angle	Python	Ligand bite angle (°)
2	Cone_Angle	Python	Ligand cone angle (°)
3	Sterimol_B1	Python	Smallest distance perpendicular to the Cu-dummy vector to the edge of the ligand (Å)
4	Sterimol_B5	Python	Largest distance perpendicular to the Cu-dummy vector to the edge of the ligand (Å)
5	Sterimol_L	Python	Distance along the Cu-dummy vector to the edge of the ligand (Å)
6	PC_Buried_Volume_3-5Å	Python	Percentage buried volume at a 3.5 Å radius (%)
7	PC_Buried_Volume_5Å	Python	Percentage buried volume at a 5 Å radius (%)
8	PC_Buried_Volume_7Å	Python	Percentage buried volume at a 7 Å radius (%)
9	SASA	Python	Solvent Accessible Surface Area (Å^2)
10	HOMO_Energy	ORCA	Energy of the HOMO (eV)
11	LUMO_Energy	ORCA	Energy of the LUMO (eV)
12	Cu-L$_x$	Python	Bond distance between Cu and ligand atom x (Å)
13	Cu-N	Python	Cu-N bond distance (Å)
14	Löwdin_Charge	ORCA	Löwdin atomic charge of the atom
15	Bonded_Valence	ORCA	Number of bonds formed by the atom
16	Atomic_Population	ORCA	Number of electrons localised on the atom
17	Bond_Order	ORCA	Number of bonds between two atoms
18	Orbital_Charge_s	ORCA	Orbital charge of the s orbital
19	Orbital_Charge_p and subshells	ORCA	Orbital charge of the p orbital and its subshells
20	Orbital_Charge_d and subshells	ORCA	Orbital charge of the d orbital and its subshells

Descriptors were automatically extracted using Python, resulting in four datasets, ***ligands_CSD_PIP_set_TSOA_NoTS***, ***ligands_CSD_PIP_set_TSSig_NoTS***, ***ligands_CSD_PYR_set_TSOA_NoTS*** and ***ligands_CSD_PYR_set_TSSig_NoTS*** containing 67 descriptors and 1683, 3708, 3990 and 5798 ligands respectively.

5.2.8.1 Initial Models

Machine learning models were built for the eight machine learning methods using the transition state independent descriptor sets. Models were assessed using an explicit training and test set. Highly correlated descriptors were removed and the data was split into training and test sets using the same ligands in each set as the transition state dependent models. The best-performing models for each dataset are shown in **Table 5.10**.

Table 5.10: Metrics of the best-performing models with all four datasets with descriptors excluding the transition states.

Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA_NoTS	SVM	0.32	6.56	79.6	55.3
ligands_CSD_PYR_set_TSOA_NoTS	SVM	0.67	5.16	68.1	42.6
ligands_CSD_PIP_set_TSSig_NoTS	SVM	0.57	3.74	83.7	61.4
ligands_CSD_PYR_set_TSSig_NoTS	ET	0.66	4.83	75.8	51.0

The best models use the SVM and ET methods. Metrics using transition state independent descriptors are similar to the transition state dependant descriptor set. Metrics for ***ligands_CSD_PYR_set_TSOA_NoTS*** are almost identical to ***ligands_CSD_PYR_set_TSOA***. The ***ligands_CSD_PIP_set_TSOA_NoTS*** performs worse using the transition state independent descriptor set with a decrease in R^2 from 0.59 to 0.32 and increase in RMSE from 5.25 to 6.56. Both **TSSig** datasets perform better with transition state independent descriptors with an improvement in all four metrics. For ***ligands_CSD_PIP_set_TSSig_NoTS*** the RMSE is within the average error in the calculated activation energies (3.9 kcal mol^{-1}). Initial models showed that it is possible to predict activation energies with a similar accuracy without using the transition state structure.

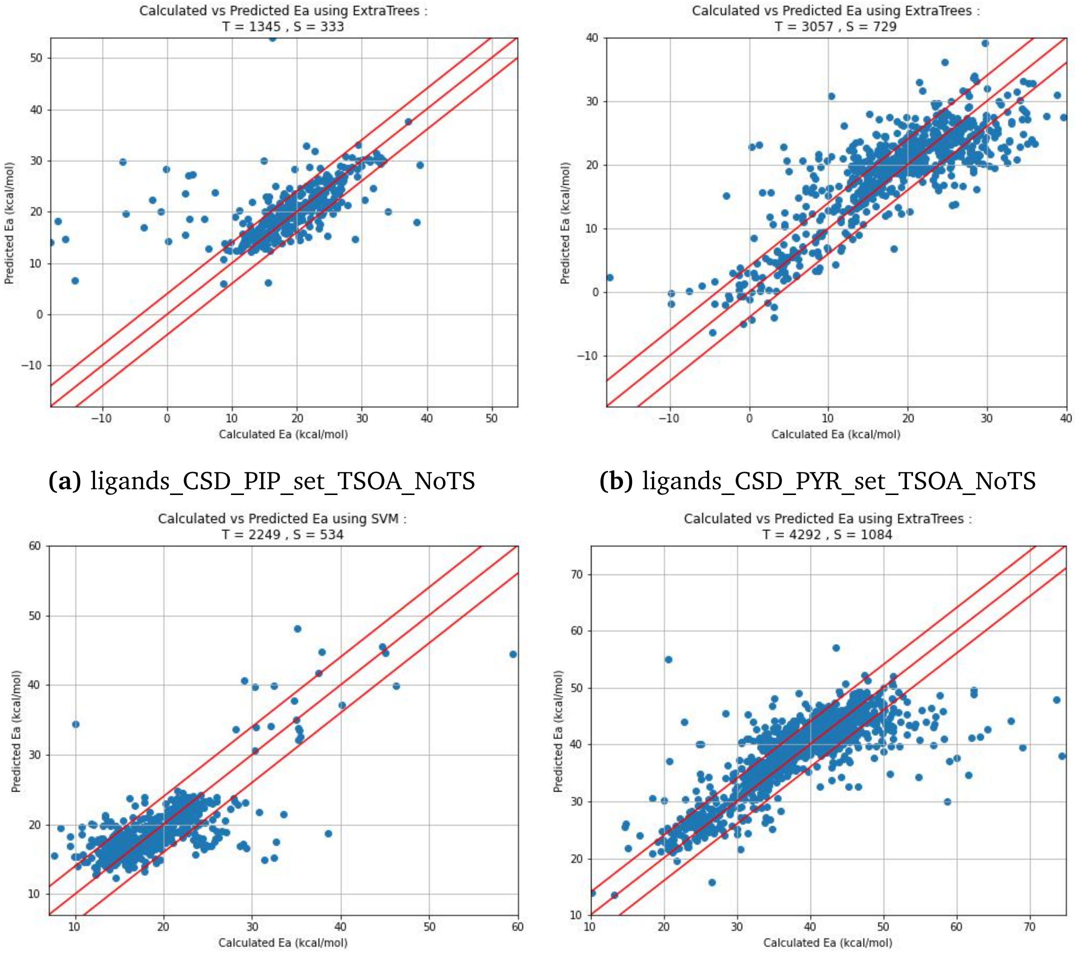
5.2.8.2 Descriptor Optimization

Permutation importance was used to identify and remove redundant and low-importance descriptors in all four datasets. Descriptors with a $mean - 2 * std > 0$ were retained resulting in descriptors sets containing 10, 27, 14 and 17 descriptors for ***ligands_CSD_-PIP_set_TSOA_NoTS***, ***ligands_CSD_PIP_set_TSSig_NoTS***, ***ligands_CSD_PYR_set_TSOA_-NoTS*** and ***ligands_CSD_PYR_set_TSSig_NoTS*** respectively. Machine learning models were rebuilt and the best-performing models for each dataset are shown in **Table 5.11**.

Table 5.11: Metrics of the best performing models, using a trimmed descriptor set, for all four datasets with descriptors excluding the transition states.

Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA_NoTS	ET	0.29	6.95	76.3	49.2
ligands_CSD_PYR_set_TSOA_NoTS	ET	0.69	4.97	66.4	41.2
ligands_CSD_PIP_set_TSSig_NoTS	SVM	0.56	3.78	82.8	58.2
ligands_CSD_PYR_set_TSSig_NoTS	ET	0.64	4.98	75.8	49.8

ExtraTrees becomes the best-performing model for the **TSOA** datasets. All metrics become slightly worse across all datasets except for ***ligands_CSD_PYR_set_TSOA_NoTS*** which shows a slight improvement in R^2 (0.67 to 0.69) and RMSE (5.16 to 4.97). For the transition state independent descriptor sets the small change in metrics suggests that there are no descriptors in the original set negatively affecting predictions. The reduction in descriptors by 40-57 only shows a small decrease in accuracy. The additional time taken to calculate the extra descriptors along with the increased time to build the models only results in a small increase in accuracy. While a smaller descriptor set is less accurate it is computationally less expensive.



(a) ligands_CSD_PIP_set_TSOA_NoTS

(b) ligands_CSD_PYR_set_TSOA_NoTS

(c) ligands_CSD_PIP_set_TSSig_NoTS

(d) ligands_CSD_PYR_set_TSSig_NoTS

Figure 5.15: Calculated activation energies vs predicted activation energies for all four TS-independent datasets using the best performing model with trimmed descriptors and optimised hyperparameters.

The transition state calculations are the most time-consuming calculations. The use of transition states independent descriptors reduces the total computational time by 44,140 single core hours of computational time for ***ligands_CSD_PIP_set*** and 48,324 single core hours for ***ligands_CSD_PYR_set***. This results in a potential decrease of 90% (**TSOA**: 67%, **TSSig**: 23%) and 84% (**TSOA**: 53%, **TSSig**: 31%) of the total computational time for the ***ligands_CSD_PIP_set*** and ***ligands_CSD_PYR_set*** respectively. The dramatic reduction in computational time not only saves on both the raw time and computational infrastructure required but also energy costs. The energy required to run the computers for sustained lengths of time, as well as the CO_2 produced, must also be considered for the sustainable use of these approaches. Eliminating as many calculations as possible is advantageous from both an infrastructure and sustainability perspective as well as making them widely accessible to the chemical community.

5.2.8.3 Important Descriptors

The importance of individual descriptors on the best-performing models for each dataset was calculated using permutation important to identify potentially important ligand properties for the active catalytic state (**CuLNu**). The top five most important descriptors for each dataset are shown in **Table 5.12**.

Table 5.12: Top five most important descriptors, by the reduction in R^2 of the ExtraTrees model for each TS independent dataset.

	ligands_CSD_PIP_set_TSOA_NoTS		***ligands_CSD_PYR_set_TSOA_NoTS***	
No.	Descriptor	Reduction in R^2	Descriptor	Reduction in R^2
1	Orbital_Charge_Cu_s	0.199 ± 0.042	Orbital_Charge_Cu_s	0.098 ± 0.008
2	Löwdin_Charge_Cu	0.041 ± 0.015	Bond_Order_Cu-N	0.075 ± 0.009
3	Orbital_Charge_Cu_p	0.013 ± 0.006	Löwdin_Charge_N	0.042 ± 0.004
4	Bite_Angle	0.011 ± 0.005	Atomic_Population_N	0.040 ± 0.005
5	Orbital_Charge_L2_dxz	0.008 ± 0.003	Cone_Angle	0.039 ± 0.013
	ligands_CSD_PIP_set_TSSig_NoTS		***ligands_CSD_PYR_set_TSSig_NoTS***	
No.	Descriptor	Reduction in R^2	Descriptor	Reduction in R^2
1	Bond_Order_Cu-N	0.136 ± 0.024	Orbital_Charge_Cu(s)	0.185 ± 0.013
2	Cone_Angle	0.033 ± 0.006	Atomic_Population_N	0.096 ± 0.011
3	LUMO_Energy	0.021 ± 0.008	Bond_Order_Cu-N	0.043 ± 0.004
4	PC_Buried_Volume_3-5A	0.018 ± 0.003	HOMO_Energy	0.032 ± 0.004
5	PC_Buried_Volume_5A	0.011 ± 0.003	LUMO_Energy	0.030 ± 0.005

For every dataset except ***ligands_CSD_PIP_set_TSSig_NoTS***, the charge on the copper *s* orbital is the most important property. The copper-nitrogen bond order is the most important property for the ***ligands_CSD_PIP_set_TSSig*** dataset. The most important descriptor for the oxidative addition transition state (**TSOA**) is the orbital charge on the copper *s* orbital. The lack of this descriptor in the ***ligands_CSD_PIP_set_TSSig_NoTS*** dataset suggests that the ability of the ligand to modulate the charge on the copper *s* orbital favours the oxidative addition transition state, with the impact of the descriptor in ***ligands_CSD_PYR_set_TSSig_NoTS*** relating to the amide nucleophile. The remaining descriptors are different between nucleophiles, for piperidine the descriptors describe the charge on the copper centre, whereas for 2-pyrrolidinone the descriptors describe the electronics at the amide nitrogen. This suggests that for the amine nucleophile, the ligand's ability to modulate the charge on the copper is important in aiding the oxidative addition of the aryl halide. The lack of amine nitrogen descriptors implies that the interaction between the copper and nucleophile is less important. For the amide, however, the ligand's ability to aid the electron donating/withdrawing ability of the copper, enabling the modulation of the electron density at the amide nitrogen is important.

For the sigma metathesis transition state, the strength of the Cu−N bond is the most important descriptor. A similar trend is seen for nucleophiles for the sigma metathesis transition state (**TSSig**). The ligand's ability to impact the electron density at the amide nitrogen is a key factor for high activity. However, for the amine, while the strength of the Cu−N bond is the most important descriptor, sterics are also shown to be important with three out of five of the descriptors describing the steric bulk of the ligand around the copper centre, albeit with relatively low importance (<0.033).

The trend in important features is consistent between the transition state dependent and transition state independent models. The oxidative addition transition state is dependent on the ligand's ability to modulate the electron density at the copper centre, aiding the oxidative addition of the aryl halide. The sigma metathesis transition state is dependent on the ligand's ability to affect the strength of the Cu−N bond. A similar trend is also seen for the different nucleophiles. For piperidine, there is no direct impact of the ligand on the electronics of the nucleophile. Whereas for the amide the ligand's ability to influence the electronics of the amide nitrogen and the amide bond is an important determinant for activity.

5.2.9 Effect of DFT Functional on Descriptors

The ability of the transition state independent descriptor sets to provide predictions similar to those of the transition state dependent descriptor sets means that the calculation of the transition states is not required for a good prediction of activity. As the transition states no longer need to be calculated the freed-up computational time could be used to generate more accurate descriptors using a higher level of theory. As seen previously electronic descriptors are more important than steric descriptors, comprising the majority of the top 5 descriptors across all datasets. To improve the electronic descriptors, several DFT methods were chosen, with differing types and amounts of HF exchange (**Table 5.13**). r^2SCAN-3c is a new improved 3c method which is better for the calculation of molecular properties,[120] PBE0 is a standard functional commonly used for transition metal complexes[125] and TPSS and TPSSh have shown to be excellent at predicting the properties of 1st-row transition metal complexes.[124]

Table 5.13: DFT methods used for benchmarking the electronic descriptors. The increase in computational time is in comparison to B97-3c for 50 ligands.

Method	Type	HF Exchange (%)	Increase in Computational Time
B97-3c	Composite GGA	0	0
r^{2}SCAN-3c	Composite meta-GGA	0	1.5-2×
PBE0	Hybrid-GGA	20	6-8×
TPSS	meta-GGA	0	2-6×
TPSSh	Hybrid-meta-GGA	10	6-8×

The correlation of each descriptor was calculated with respect to the value calculated at the DLNPO-CCSD(T)/def2-TZVPP level of theory. DLPNO-CCSD(T) is a pure wavefunction method and therefore is not influenced by experimental data, therefore, obtained electronic descriptor values are the best values possible with currently available methods. PBE0, TPSS and TPSSh calculations use the def2-TZVP basis set with D4 dispersion correction. Electronic descriptors were calculated for 50 randomly selected ligands. All ligands were present in all four datasets and span the full range of activation energies. The resulting correlations are shown in **Table 5.14**.

Table 5.14: Correlation of electronic descriptors between B97-3c, r^{2}SCAN-3c, PBE0, TPSS and TPSSh with DLPNO- CCSD(T)/def2-TZVPP. Red is a low correlation, green is a high correlation.

Method	HOMO Energy (eV)	LUMO Energy (eV)	Löwdin Charge (Cu)	Löwdin Charge (N)	Löwdin Charge (L1)	Löwdin Charge (L2)	Bonded Valence (Cu)	Bonded Valence (N)
B97-3c	0.48	0.85	0.01	0.71	0.78	0.67	0.5	0.21
PBE0	0.75	0.9	0.71	0.91	0.78	0.94	0.54	0.67
r2SCAN-3c	0.49	0.85	−0.07	0.69	0.39	0.6	0.39	0.45
TPSS	0.52	0.85	0.5	0.87	0.71	0.91	0.38	0.6
TPSSh	0.63	0.87	0.58	0.89	0.75	0.92	0.46	0.64

Method	Bonded Valence (L1)	Bonded Valence (L2)	Atomic Population (Cu)	Atomic Population (N)	Atomic Population (L1)	Atomic Population (L2)	Bond Order (Cu-N)	Bond Order (Cu-L1)
B97-3c	0.98	0.99	0.2	0.43	1	0.99	0.84	0.9
PBE0	1	1	0.76	0.71	1	1	0.85	0.96
r2SCAN-3c	0.98	0.95	0.08	0.36	1	0.99	0.79	0.79
TPSS	0.99	0.99	0.76	0.51	1	1	0.83	0.95
TPSSh	1	1	0.77	0.6	1	1	0.84	0.96

Method	Orbital Charge N(s)	Orbital Charge N(p)	Orbital Charge Cu(s)	Orbital Charge Cu(p)	Orbital Charge Cu(d)	Orbital Charge Cu(dxz)	Orbital Charge Cu(dyz)	Orbital Charge Cu(dxy)
B97-3c	0.78	0.69	0.89	0.91	0.31	0.47	0.3	0.1
PBE0	0.94	0.89	0.97	0.97	0.82	0.78	0.65	0.71
r2SCAN-3c	0.92	0.78	0.92	0.98	0.26	0.45	0.26	0.08
TPSS	0.94	0.82	0.95	0.96	0.69	0.67	0.49	0.49
TPSSh	0.94	0.85	0.96	0.96	0.74	0.71	0.55	0.58

Method	Orbital Charge Cu(dz2)	Orbital Charge Cu(dx2y2)	Orbital Charge L1(s)	Orbital Charge L1(p)	Orbital Charge L1(d)	Orbital Charge L2(s)	Orbital Charge L2(p)	Orbital Charge L2(d)
B97-3c	0.71	0.04	0.99	1	−0.86	1	1	−0.89
PBE0	0.92	0.83	1	1	1	1	1	1
r2SCAN-3c	0.65	−0.03	0.99	1	0.94	0.99	0.99	0.98
TPSS	0.82	0.48	1	1	1	1	1	1
TPSSh	0.87	0.64	1	1	1	1	1	1

B97-3c shows a poor correlation for the Löwdin charge on copper as well as all *d* orbitals. All other methods correlate well with DLPNO-CCSD(T)/def2-TZVPP for orbital charges except copper the copper *d* orbitals, descriptors localised on the ligand coordinating atoms, the LUMO energy and bond orders. For Löwdin charges, descriptors localised on copper and the nucleophile nitrogen, as well as copper *d* orbitals, correlate with the amount of HF exchange in the functional. A higher percentage of HF exchange has a better correlation with DLPNO-CCSD(T)/def2-TZVPP. This suggests that a large amount of HF exchange is required to correctly describe the bonding between the copper centre and the nucleophile nitrogen atom. All other descriptors can be sufficiently described without the inclusion of HF exchange.

5.2.9.1 Effect of DFT Functional on Predicted Activation Energies

Electronic descriptors are calculated from the energy calculation in the workflow therefore, as the calculation needs to be redone for each method, activation energies can also be calculated for these methods. The activation energies for the same 50 ligands were compared with DLPNO-CCSD(T)/def2-TZVPP calculated values. Root mean squared errors were calculated for both the raw values and values scaled to DLPNO-CCSD(T)/def2-TZVPP using the equation of the line.

Table 5.15: Correlation of activation energies between B97-3c and 4 other DFT methods and DLPNO-CCSD(T) for 50 ligands. RMSE_Actual is the RMSE of the raw value of the activation energy compared to DLPNO-CCSD(T)/def2-TZVPP. RMSE_Scaled is the RMSE of the scaled activation energy using the equation of the line to convert to a DLPNO-CCSD(T)/def2-TZVPP energy.

	TSOA Activation Energy			TSSig Activation Energy		
	R^2	RMSE_Actual	RMSE_Scaled	R^2	RMSE_Actual	RMSE_Scaled
B97-3c	0.91	5.97	3.78	0.87	8.96	3.46
r^2SCAN-3c	0.90	8.87	4.07	0.90	7.20	3.05
PBE0	0.96	4.67	2.54	0.97	4.34	1.65
TPSS	0.88	10.01	4.30	0.88	8.87	3.33
TPSSh	0.92	8.40	3.67	0.93	7.14	2.57

All methods correlate reasonably well with DLPNO-CCSD(T)/def2-TZVPP, with R^2 >0.87. RMSE decreases with an increasing percentage of HF exchange in the functional. Surprisingly B97-3c correlates very well for **TSOA** compared with the other high HF exchange methods. Errors are higher for **TSOA** across all functionals. PBE0 is the best-performing functional, out of the selected functionals, for the calculation of activation energies. Descriptors calculated at the PBE0/def2-TZVP level of theory provide the most accurate electronic descriptors and activation energies for the same computational cost of the transition state calculations they are replacing.

5.2.9.2 Transition State Independent Models

Activation energies and electronic descriptors were recalculated for the structures in the machine-learning datasets at the TPSS/def2-TZVP and PBE0/def2-TZVP levels of theory, using D4 dispersion correction, to generate four ***_TPSS** and four ***_PBE0** datasets respectively. TPSS was chosen as a middle ground between accuracy and computational time and PBE0 was chosen as the most accurate method within the computational time constraints.

Machine learning models were built using the same eight machine learning methods. Models used the same explicit training and test sets, using the same ligands in each set. The best-performing models for each dataset are summarised in **Table 5.16** and **Table 5.17**.

Table 5.16: Metrics of the best performing models, with TS independent descriptors, for all four datasets with descriptors and activation energies calculated at the PBE0/def2-TZVP level of theory.

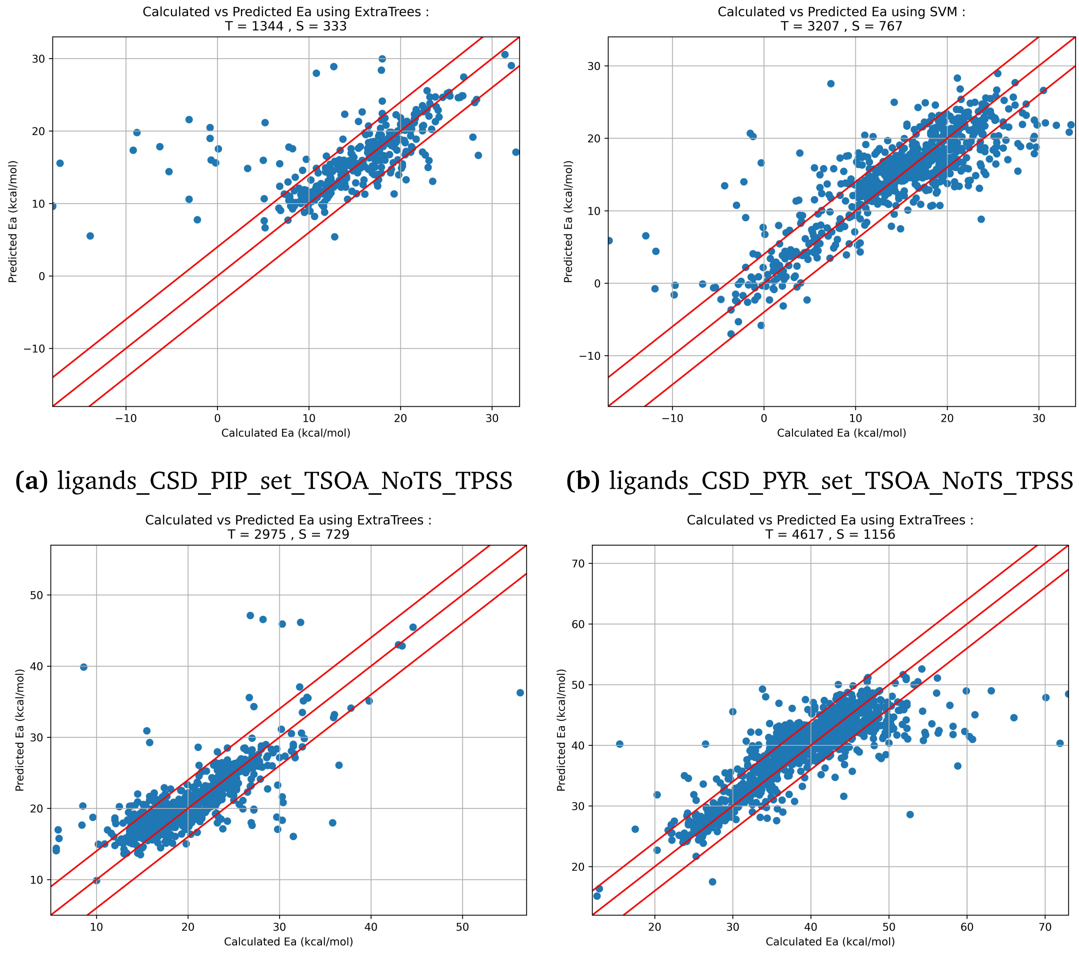
Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA_NoTS_PBE0	SVM	0.40	6.11	82.3	58.6
ligands_CSD_PYR_set_TSOA_NoTS_PBE0	SVM	0.71	4.59	72.6	47.3
ligands_CSD_PIP_set_TSSig_NoTS_PBE0	SVM	0.69	3.78	84.6	62.0
ligands_CSD_PYR_set_TSSig_NoTS_PBE0	ET	0.68	4.66	78.0	54.7

In general, DFT descriptors gave better predictions, especially for the ***ligands_CSD_PYR_-set_TSOA_NoTS_PBE0*** dataset with an improvement in RMSE of 0.5 kcal mol^{-1}. Performance metrics were uniformly improved across both DFT methods.

Table 5.17: Metrics of the best performing models, with TS independent descriptors, for all four datasets with descriptors and activation energies calculated at the TPSS/def2-TZVP level of theory.

Dataset	Best Method	R^2	RMSE	% within 4.0	% within 2.0
ligands_CSD_PIP_set_TSOA_NoTS_TPSS	ET	0.34	6.03	80.5	58.0
ligands_CSD_PYR_set_TSOA_NoTS_TPSS	SVM	0.69	4.27	75.4	52.3
ligands_CSD_PIP_set_TSSig_NoTS_TPSS	ET	0.62	3.46	87.8	67.8
ligands_CSD_PYR_set_TSSig_NoTS_TPSS	ET	0.70	4.09	80.4	55.6

Surprisingly, the TPSS models performed better than PBE0 with better RMSE metrics across all four datasets. The ***ligands_CSD_PIP_set_TSSig_NoTS_TPSS*** dataset is well within the error of the calculated activation energies (3.9 kcal mol^{-1}). All datasets have at least 75% of the predictions within 4 kcal mol^{-1}. DFT can be used to provide more accurate predictions at the cost of increased computational time and resources to both generate the initial model and descriptors.



(a) ligands_CSD_PIP_set_TSOA_NoTS_TPSS

(b) ligands_CSD_PYR_set_TSOA_NoTS_TPSS

(c) ligands_CSD_PIP_set_TSSig_NoTS_TPSS

(d) ligands_CSD_PYR_set_TSSig_NoTS_TPSS

Figure 5.16: TPSS/def2-TZVP calculated activation energies vs predicted activation energies for all four TS-independent datasets using the best performing model. Descriptors were calculated at the TPSS/def2-TZVP level of theory.

5.3 Conclusions

In conclusion, machine learning models were built and tested using eight machine learning methods. The models proved effective at identifying incorrect structures that are not picked up during structure validation in the high-throughput computational workflow as they appear as significant outliers in the models.

Models were analysed with the following conclusions: tree-based methods (RF, ET and Bagging) and SVM are superior for the prediction of activation energies; linear methods and neural networks perform poorly across all datasets, and under-represented interactions such as hydrogen bonding and poorly described with the descriptor set and lead to increased error. Smaller datasets lead to poorer metrics with near-identical descriptor sets due to the lack of data required to sufficiently train the models.

Analysis of descriptor importance can provide insight into the underlying chemistry of more complex, less understood mechanisms providing possible routes for further exploration of ligand properties and their effect on activity.

Transition state structures are not required to produce accurate machine-learning models. Models not including descriptors derived from the transition state can provide equal if not better predictions of activation energies. The lack of need for a transition state structure allows for the computational budget to be used elsewhere. DFT-derived electronic descriptors further improved models with prediction errors close to the errors in the dataset.

5.4 Methodology

Example machine learning models, an hyperparameter optimisation script, an example feature importance script and datasets can be found at `https://github.com/MarcS18/Thesis_ESI`.

5.4.1 Machine Learning Overview

Eight machine learning models were employed; Multiple Linear Regression (MLR), Gaussian Process Regression (GP), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Partial Least Squares (PLS), Random Forest (RF), ExtraTrees (ET) and Bagging (Bag). Default parameters were used with the following exceptions: for GP only the Matern. RBF and RationalQuadratic kernel were used; for ANN, `n_nodes` (number of nodes in the hidden layers) was optimised with the number of hidden layers varied; for SVM the radial basis function (RBF) kernel was used with C, epsilon and gamma being optimised; for PLS, `n_components` (number of components to retain after dimension reduction) was optimised; and for RF, ET and Bag, `n_estimators` (number of trees) and `max_depth` was optimised. Machine learning was performed in Python 3 with the scikit-learn module. Prior to machine learning, all descriptors were scaled between 0 and 1 using scikit-learn's `StandardScaler()` method. The modules and parameters used in this chapter are summarised in **Table 5.18**.

Table 5.18: The scikit-learn methods and parameters used in this chapter.

Method	scikit-learn method	Parameters
MLR	`linear_model.LinearRegression()`	Default
GP	`gaussian_process.GaussianProcessRegressor()`	`kernel` optimised
ANN	`neural_network.MLPRegressor()`	`n_estimators` optimised
SVM	`svm.SVR()`	`kernel=rbf`, C, epsilon and gamma optimised
PLS	`cross_decomposition.PLSRegression()`	`n_components` optimised
RF	`ensemble.RandomForestRegressor()`	`n_estimators` and `max_depth` optimised
ET	`ensemble.ExtraTreesRegressor()`	`n_estimators` and `max_depth` optimised
Bag	`ensemble.BaggingRegressor()`	`n_estimators` optimised

5.4.2 Parameter Optimization

Parameters for all models were optimised using the Optuna python package.[230] All methods used the same seed to ensure repeatability between runs. All parameters of interest were optimised within a specified range using the `study.optimise()` function with the method set to optimise to a maximum for the coefficient of determination (R^2). Note that only the Coefficient of Determination is minimised and based entirely on optimizing the parameters for the training data, thus can be worse on the new test data. If the best values for the coefficient of determination were obtained close to the boundary of the range of tested values for a specific parameter the range was expanded and the optimisation was rerun. The optimised parameters were retained for use in future models.

5.4.2.1 Linear Regression

For linear regression as no parameters need to be optimised the coefficient of determination was taken from 1 trial of the optimiser.

5.4.2.2 Gaussian Process

Three kernels are tested for Gaussian Process models: Matern, RBF and RationalQuadratic. Kernels were tested over k trials where k is the number of K-Folds used. The best kernel was saved for later models.

5.4.2.3 Partial Least Squares (PLS)

For partial least squares the number of components, `n_components`, was varied between 1 and the maximum number of features (N_f), retaining the best value over N_f trials.

5.4.2.4 Support Vector Machine (SVM)

Parameters C, epsilon and gamma were optimised over 50 trials, retaining the best value. C was varied from 1×10^{-6} to 1×10^{4}, epsilon from 1×10^{-10} to 1×10^{2} and gamma from 1×10^{-4} to 1×10^{3} in log scale.

5.4.2.5 Random Forest (RF), ExtraTrees (ET) and Bagging (Bag)

The number of trees (`n_estimators`) was varied between 10 and 1000 over 50 trials, with 10-fold cross-validation of the training set. For RandomForest and ExtraTrees, the maximum depth of the tree (`max_depth`) was varied between 2-100 with the best values retained.

5.4.2.6 Artificial Neural Network (ANN)

For ANN, two hidden layers were used. The number of nodes in this layer (`n_estimators`) was varied between 10 and 1000 over 50 trials, with a maximum number of 800 iterations for convergence, and the best value retained.

5.4.3 Machine Learning Model Evaluation and Metrics

5.4.3.1 Creating training and test sets

Datasets were split into training and test sets by binning the data in intervals of 1 kcal mol^{-1}. A proportional amount of data was taken from each bin to form a training set (∼80% of the data) and a test set (∼20% of the data). Each model was trained on the same training set and tested on the same unseen test set.

5.4.3.2 K-fold Cross validation

Performance metrics are obtained by splitting the data into k groups using the scikit-learn's KFolds method ensuring the dataset was shuffled before splitting. Each group is used as a test set and the remaining $k-1$ groups are used as the training set. After each group, the performance metrics are stored and the model is discarded. K-fold cross-validation was used due to the small nature of the dataset ensuring efficient use of the data. Every data point is used at least once in the test set, avoiding chance bias possible in a static train/test split. All stated uses of K-fold cross-validation use 10 folds.

5.4.3.3 Metrics

The following metrics were used to evaluate how well the predictions compare to the calculated values of $\Delta G^{\ddagger}$. Pearson's R^2 is a measure of correlation. Root Mean Square Error (RMSE) is a measure of the error associated between each prediction and actual value. These are defined in **Equation 5.1** and **Equation 5.2** respectively.

$$R^2 = \left(\frac{\sum(x_{pred} - \overline{x})(y_{pred} - \overline{y})}{\sqrt{\sum(x_{pred} - \overline{x})^2 - \sum(y_{pred} - \overline{y})^2}} \right)^2 \tag{5.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5.2}$$

where $\overline{x}$ and $\overline{y}$ are the mean values of x and y. x_{pred} and y_{pred} are the predicted values of x and y. n is the number of values and $\hat{y}_i$ is the predicted value and y_i is the actual value. In addition, %± 4.0 kcal mol^{-1} and %± 2.0 kcal mol^{-1}, the percentage of values within the error and half the error of the computational values were also used as metrics.

5.4.3.4 Prediction Analysis

The correlation between descriptors was calculated using the pandas module `corr()` function in Python and plotted as a matrix coloured to R^2. Outliers were defined as having a prediction error of ±1.5× the RMSE. The scikit learn `permutation_importance()` method was used to determine the weighting of each descriptor for the RandomForest and ExtraTrees models. Descriptors were scored using the R^2, neg_mean_abolsulte_percentage_error and neg_mean_squared_error metrics. Descriptors were retained with an importance of:

$$mean - 2 * std > 0 \tag{5.3}$$

Importances are shown as mean±std.

5.A Appendix

5.A.1 Descriptor vs Activation Energy Correlations

Table 5.A.1: Correlation (R^2) between each descriptor and activation energy for every descriptor in the ***ligands_CSD_PIP_set_TSOA***, ***ligands_CSD_PYR_set_TSOA***, ***ligands_CSD_-PIP_set_TSSig*** and ***ligands_CSD_PYR_set_TSSig*** datasets.

Descriptor	R^2			
	TSOA_PIP	TSOA_PYR	TSSig_PIP	TSSig_PYR
Bite Angle	−0.03	−0.02	−0.29	−0.10
Change in Bite Angle	−0.03	0.11	−0.06	0.06
Cone Angle	0.11	0.23	−0.22	−0.02
Sterimol B1	0.02	−0.01	0.00	−0.06
Sterimol B5	−0.05	0.03	−0.11	−0.01
Sterimol L	−0.09	−0.13	0.05	−0.11
PC_Buried_Volume_3-5A	0.16	0.26	−0.17	0.07
PC_Buried_Volume_5A	0.12	0.21	−0.18	0.03
PC_Buried_Volume_7A	0.04	0.09	−0.16	−0.03
SASA	−0.06	−0.05	−0.09	−0.09
HOMO Energy	0.29	0.29	0.05	0.47
LUMO Energy	0.17	0.24	0.10	0.46
Cu-L1	−0.05	0.05	0.11	0.03
Cu-L2	−0.03	0.07	0.07	0.04
D_Cu-L1	0.05	−0.08	0.08	−0.04
D_Cu-L2	0.00	−0.14	0.03	−0.02
Cu-I	0.13	0.25	−0.24	0.17

Continuation of Table 5.A.1

Descriptor	TSOA_PIP	TSOA_PYR	TSSig_PIP	TSSig_PYR
Cu-N	-	-	−0.21	0.19
Cu-C	0.32	0.16	−0.27	0.38
Cu-O	-	-	-	−0.04
C-I	−0.29	−0.18	0.12	−0.18
Amide C-O	-	-	-	0.27
Amide C-N	-	-	-	−0.32
N-Cu-I	-	-	0.27	−0.23
Cu-I-C	0.35	0.15	0.07	0.33
I-C-N	-	-	−0.23	0.04
C-N-Cu	-	-	−0.29	−0.05
I-C-Cu	0.14	0.20	-	-
C-Cu-I	−0.31	−0.21	-	-
Amide O-C-N	-	-	-	0.29
Löwdin Charge (Cu)	−0.14	−0.14	−0.07	−0.05
Löwdin Charge (N)	-	-	0.00	−0.28
Löwdin Charge (C)	−0.29	−0.20	0.05	−0.32
Löwdin Charge (I)	0.30	0.17	0.10	0.03
Löwdin Charge (L1)	0.07	0.03	−0.01	−0.14
Löwdin Charge (L2)	0.09	0.02	−0.04	−0.16
Löwdin Charge (Amide C)	-	-	-	−0.18
Löwdin Charge (Amide O)	-	-	-	−0.28
Bonded Valence (Cu)	−0.11	−0.03	0.12	0.00
Bonded Valence (N)	-	-	0.14	−0.29
Bonded Valence (C)	0.01	0.03	−0.06	0.23
Bonded Valence (I)	0.33	0.14	0.24	0.01
Bonded Valence (L1)	0.07	0.07	−0.04	−0.09
Bonded Valence (L2)	0.10	0.06	−0.06	−0.1
Bonded Valence (Amide C)	-	-	-	0.06
Bonded Valence (Amide O)	-	-	-	−0.14
Atomic Population (Cu)	−0.11	0.07	0.37	0.38
Atomic Population (N)	-	-	−0.10	−0.20
Atomic Population (C)	0.16	0.16	0.11	−0.12
Atomic Population (I)	−0.19	0.05	−0.04	0.10
Atomic Population (L1)	−0.05	−0.03	−0.01	0.04
Atomic Population (L2)	−0.07	−0.08	0.05	0.05
Atomic Population (Amide C)	-	-	-	−0.05
Atomic Population (Amide O)	-	-	-	0.13
Bond Order (Cu-I)	−0.28	−0.27	0.27	−0.03
Bond Order (Cu-C)	−0.34	−0.15	0.03	0.06
Bond Order (C-I)	0.25	0.21	−0.20	0.21
Bond Order (Cu-N)	-	-	0.05	−0.30
Bond Order (Cu-L1)	0.04	0.06	−0.07	0.02
Bond Order (Cu-L2)	0.05	0.00	−0.01	0.00
Bond Order (Amide C-O)	-	-	-	−0.16
Bond Order (Amide C-N)	-	-	-	0.23
Orbital Charge C(s)	0.31	0.10	0.13	0.11
Orbital Charge C(p)	0.26	0.20	−0.06	0.28
Orbital Charge C(pz)	0.04	0.17	−0.05	−0.02
Orbital Charge C(px)	−0.02	0.13	0.03	−0.11

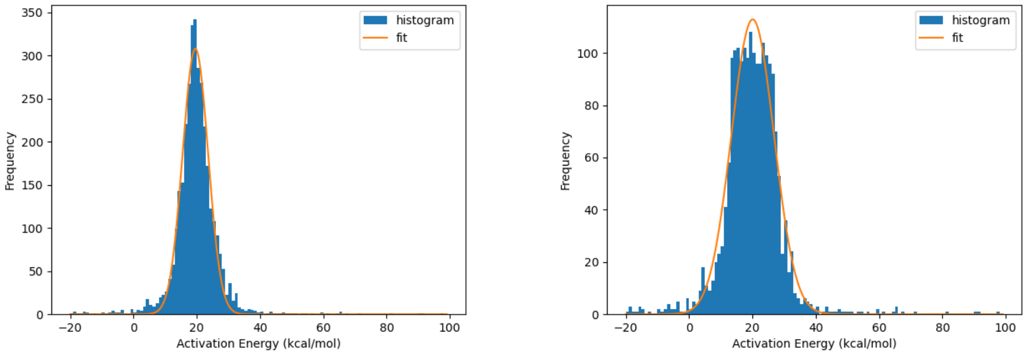
Continuation of Table 5.A.1

Descriptor	TSOA_PIP	TSOA_PYR	TSSig_PIP	TSSig_PYR
Orbital Charge C(py)	0.12	−0.08	−0.03	0.17
Orbital Charge N(s)	-	-	−0.17	0.13
Orbital Charge N(p)	-	-	0.03	0.21
Orbital Charge N(pz)	-	-	0.06	0.16
Orbital Charge N(px)	-	-	−0.06	0.04
Orbital Charge N(py)	-	-	0.01	−0.17
Orbital Charge I(s)	−0.39	−0.25	−0.23	−0.26
Orbital Charge I(p)	−0.33	−0.20	−0.17	−0.06
Orbital Charge I(pz)	−0.14	0.11	0.02	0.11
Orbital Charge I(px)	−0.15	−0.11	−0.08	−0.13
Orbital Charge I(py)	0.03	−0.13	−0.03	−0.02
Orbital Charge I(d)	0.48	0.33	0.21	0.40
Orbital Charge I(dxz)	0.23	0.06	0.06	0.09
Orbital Charge I(dyz)	0.16	0.21	0.08	0.06
Orbital Charge I(dxy)	0.20	0.19	0.17	0.25
Orbital Charge I(dz2)	0.25	0.10	0.01	0.06
Orbital Charge I(dx2y2)	0.04	0.26	0.10	0.32
Orbital Charge Cu(s)	0.11	0.00	0.01	0.00
Orbital Charge Cu(p)	−0.08	−0.02	0.05	−0.11
Orbital Charge Cu(pz)	−0.01	−0.22	0.18	−0.20
Orbital Charge Cu(px)	−0.15	−0.07	−0.09	0.06
Orbital Charge Cu(py)	0.01	0.25	−0.02	−0.02
Orbital Charge Cu(d)	0.27	0.24	0.05	0.19
Orbital Charge Cu(dxz)	0.16	0.04	0.05	0.11
Orbital Charge Cu(dyz)	0.04	0.06	−0.08	0.31
Orbital Charge Cu(dxy)	−0.03	−0.11	−0.03	0.05
Orbital Charge Cu(dz2)	0.04	0.18	0.09	−0.22
Orbital Charge Cu(dx2y2)	0.05	0.06	0.10	−0.16
Orbital Charge L1(s)	−0.03	−0.02	0.00	0.04
Orbital Charge L1(p)	−0.03	0.00	−0.02	0.06
Orbital Charge L1(pz)	−0.01	0.05	0.00	0.13
Orbital Charge L1(px)	0.03	0.02	−0.07	0.01
Orbital Charge L1(py)	−0.09	−0.06	0.05	0.05
Orbital Charge L1(d)	−0.09	−0.08	−0.02	0.01
Orbital Charge L1(dxz)	−0.11	−0.06	0.05	0.11
Orbital Charge L1(dyz)	−0.05	0.00	−0.08	0.02
Orbital Charge L1(dxy)	−0.08	−0.12	−0.03	−0.04
Orbital Charge L1(dz2)	−0.10	−0.01	0.00	0.08
Orbital Charge L1(dx2y2)	−0.08	−0.11	−0.01	−0.05
Orbital Charge L2(s)	−0.05	−0.05	0.06	0.05
Orbital Charge L2(p)	−0.06	−0.04	0.04	0.08
Orbital Charge L2(pz)	−0.04	0.01	0.04	0.14
Orbital Charge L2(px)	0.01	−0.04	−0.02	0.03
Orbital Charge L2(py)	−0.10	−0.07	0.12	0.05
Orbital Charge L2(d)	−0.08	−0.09	0.00	−0.01
Orbital Charge L2(dxz)	−0.09	−0.08	0.08	0.07
Orbital Charge L2(dyz)	−0.06	−0.04	−0.07	−0.01
Orbital Charge L2(dxy)	−0.08	−0.11	−0.01	−0.06
Orbital Charge L2(dz2)	−0.08	−0.06	0.03	0.06

Continuation of Table 5.A.1

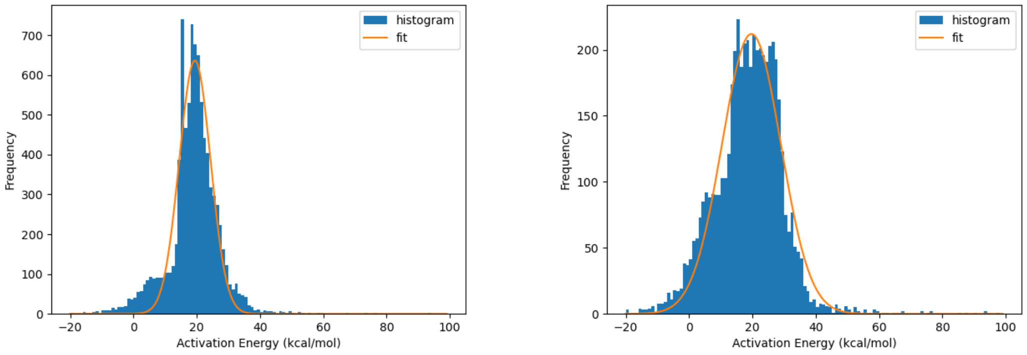
Descriptor	TSOA_PIP	TSOA_PYR	TSSig_PIP	TSSig_PYR
Orbital Charge L2(dx2y2)	−0.07	−0.10	0.00	−0.06
Orbital Charge Amide C(s)	-	-	-	−0.23
Orbital Charge Amide C(p)	-	-	-	0.34
Orbital Charge Amide C(pz)	-	-	-	−0.11
Orbital Charge Amide C(px)	-	-	-	−0.03
Orbital Charge Amide C(py)	-	-	-	0.16
Orbital Charge Amide O(s)	-	-	-	−0.10
Orbital Charge Amide O(p)	-	-	-	0.27
Orbital Charge Amide O(pz)	-	-	-	0.12
Orbital Charge Amide O(px)	-	-	-	−0.14
Orbital Charge Amide O(py)	-	-	-	0.19
Magnitude of the Imaginary Frequency	0.00	0.16	0.02	0.33

5.A.2 Activation Energy Distributions



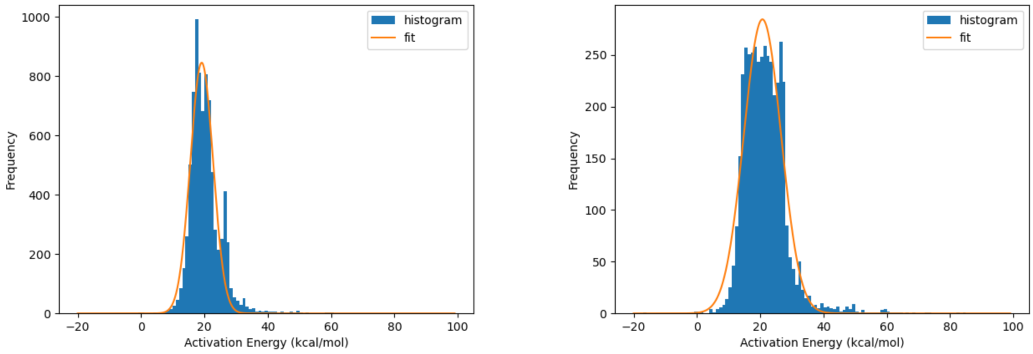
(a) Before Trimming **(b)** After Trimming

Figure 5.A.1: Activation Energy distributions before and after trimming for the ***ligands_-CSD_PIP_set_TSOA*** dataset.



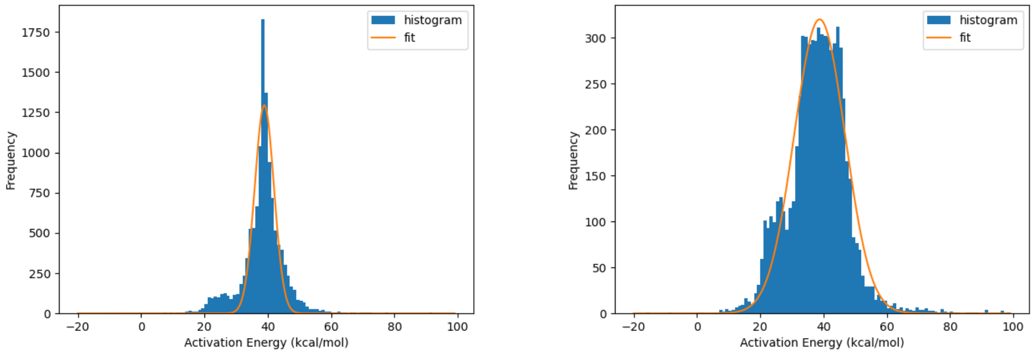
(a) Before Trimming **(b)** After Trimming

Figure 5.A.2: Activation Energy distributions before and after trimming for the ***ligands_-CSD_PYR_set_TSOA*** dataset.



(a) Before Trimming **(b)** After Trimming

Figure 5.A.3: Activation Energy distributions before and after trimming for the ***ligands_-CSD_PIP_set_TSSig*** dataset.



(a) Before Trimming **(b)** After Trimming

Figure 5.A.4: Activation Energy distributions before and after trimming for the ***ligands_-CSD_PYR_set_TSSig*** dataset.

5.A.3 Machine Learning Models

For all graphs, T is the amount of training data and S is the amount of test data. Red lines are parity, $+3.9\ \mathrm{kcal\,mol^{-1}}$ and $-3.9\ \mathrm{kcal\,mol^{-1}}$. Eight machine learning models were tested, MLR = Multiple Linear Regression, GPR = Gaussian Process Regression, ANN = Artificial Neural Network, SVM = Support Vector Machine, PLS = Partial Least Squares, RF = Random Forest, ExtraTrees and Bagging.

5.A.4 Initial Models

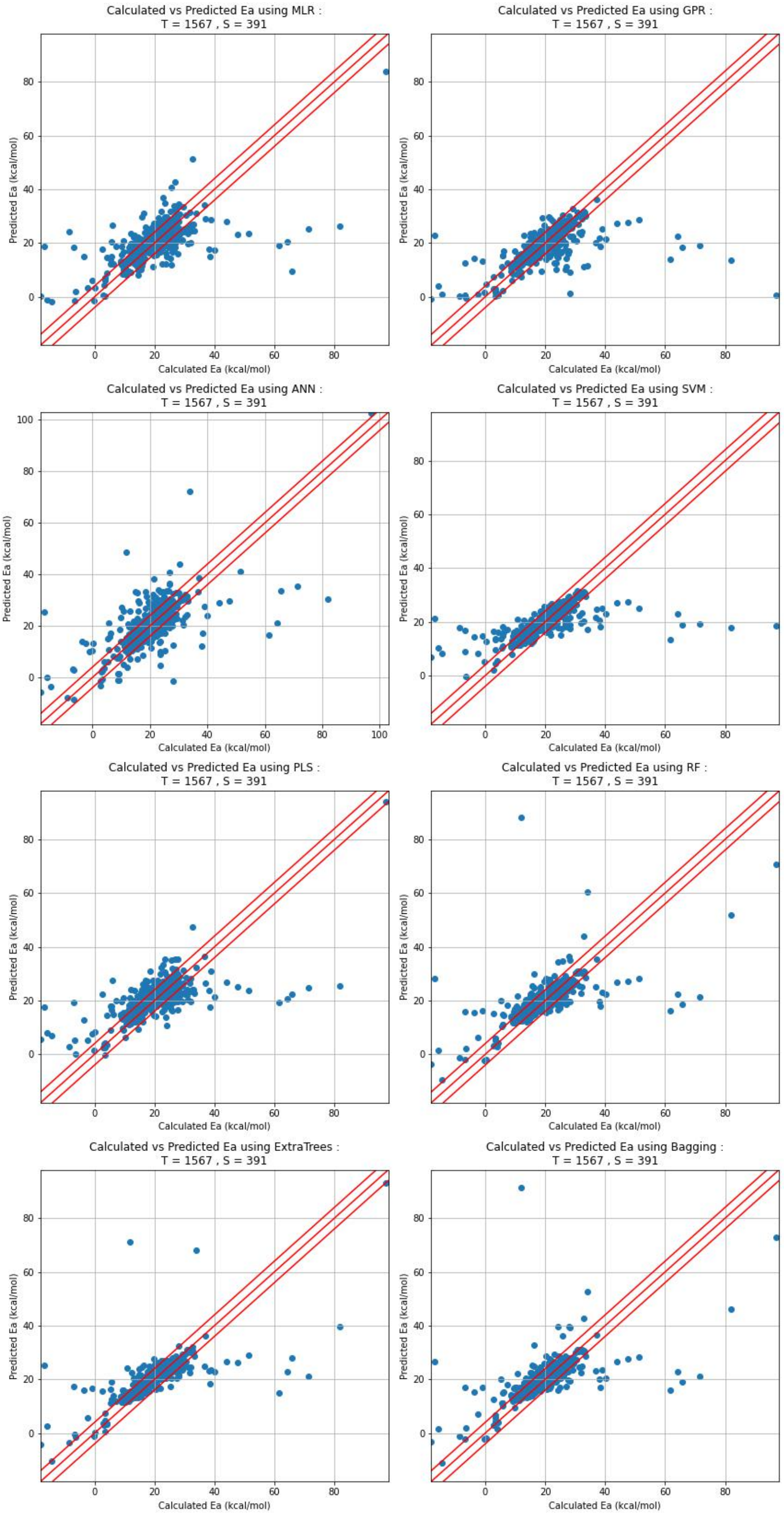
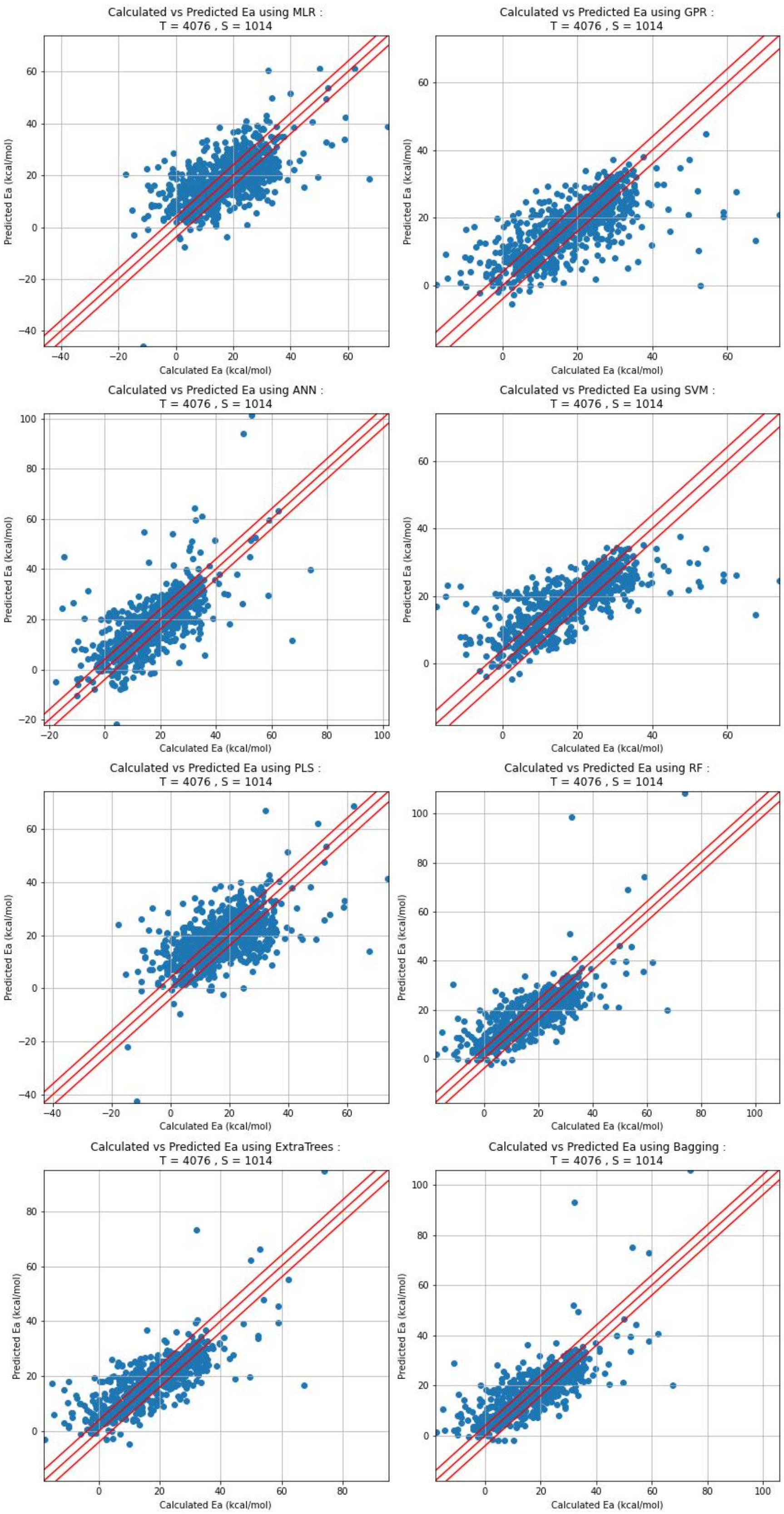


Figure 5.A.5: Initial machine learning models for the ***ligands_CSD_PIP_set_TSOA*** dataset.



Calculated vs Predicted Ea using MLR : T = 4076 , S = 1014

Calculated vs Predicted Ea using GPR : T = 4076 , S = 1014

Calculated vs Predicted Ea using ANN : T = 4076 , S = 1014

Calculated vs Predicted Ea using SVM : T = 4076 , S = 1014

Calculated vs Predicted Ea using PLS : T = 4076 , S = 1014

Calculated vs Predicted Ea using RF : T = 4076 , S = 1014

Calculated vs Predicted Ea using ExtraTrees : T = 4076 , S = 1014

Calculated vs Predicted Ea using Bagging : T = 4076 , S = 1014

Figure 5.A.6: Initial machine learning models for the ***ligands_CSD_PYR_set_TSOA*** dataset.

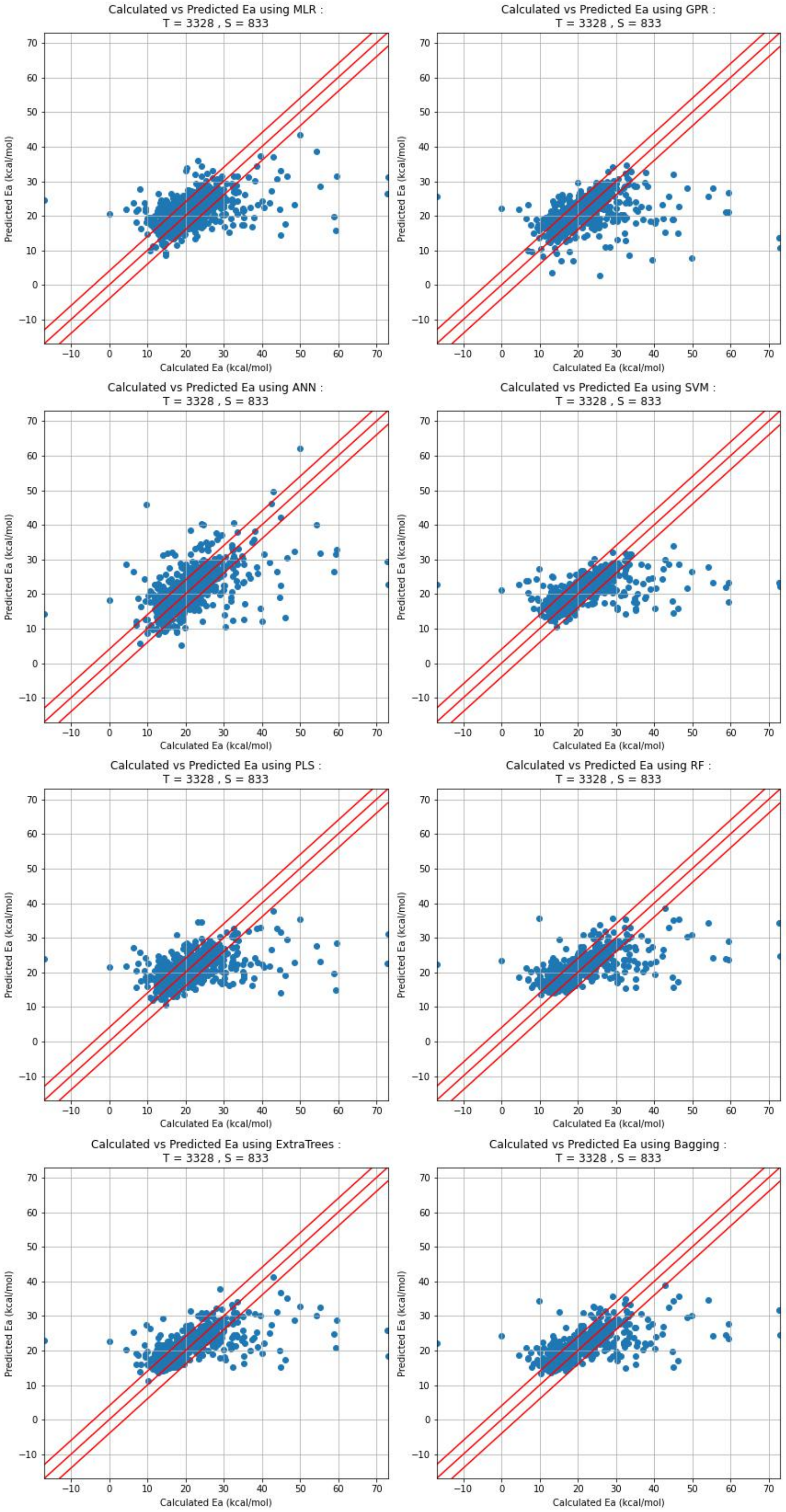
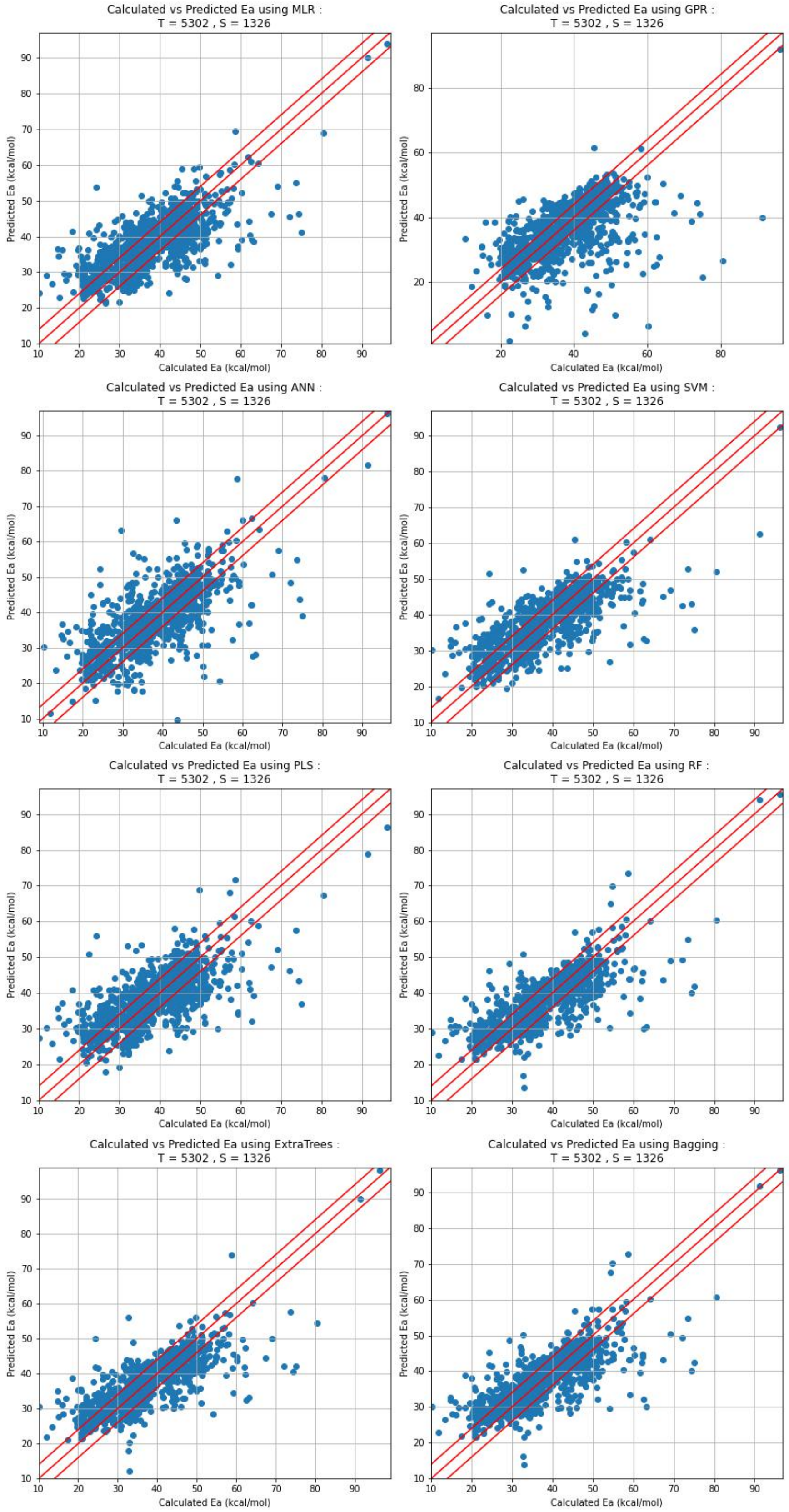


Figure 5.A.7: Initial machine learning models for the ***ligands_CSD_PI_set_TSSig*** dataset.



Calculated vs Predicted Ea using MLR : T = 5302 , S = 1326

Calculated vs Predicted Ea using GPR : T = 5302 , S = 1326

Calculated vs Predicted Ea using ANN : T = 5302 , S = 1326

Calculated vs Predicted Ea using SVM : T = 5302 , S = 1326

Calculated vs Predicted Ea using PLS : T = 5302 , S = 1326

Calculated vs Predicted Ea using RF : T = 5302 , S = 1326

Calculated vs Predicted Ea using ExtraTrees : T = 5302 , S = 1326

Calculated vs Predicted Ea using Bagging : T = 5302 , S = 1326

Figure 5.A.8: Initial machine learning models for the ***ligands_CSD_PYR_set_TSSig*** dataset.

Table 5.A.2: Initial machine learning metrics for the ***ligands_CSD_PIP_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.39	8.64	58.8	32.5
GPR	0.25	9.70	72.9	49.9
ANN	0.45	8.55	61.4	40.9
SVM	0.33	9.14	74.4	55.0
PLS	0.44	8.31	60.1	35.3
RF	0.43	8.44	74.2	53.2
ExtraTrees	0.49	7.90	75.5	54.0
Bagging	0.41	8.56	73.2	52.9

Table 5.A.3: Initial machine learning metrics for the ***ligands_CSD_PYR_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.37	8.71	43.6	24.8
GPR	0.51	7.48	63.7	44.1
ANN	0.52	7.99	57.0	34.7
SVM	0.55	7.18	64.0	42.6
PLS	0.34	8.92	43.6	24.0
RF	0.60	6.76	64.6	42.9
ExtraTrees	0.65	6.32	66.1	45.8
Bagging	0.61	6.72	64.3	43.8

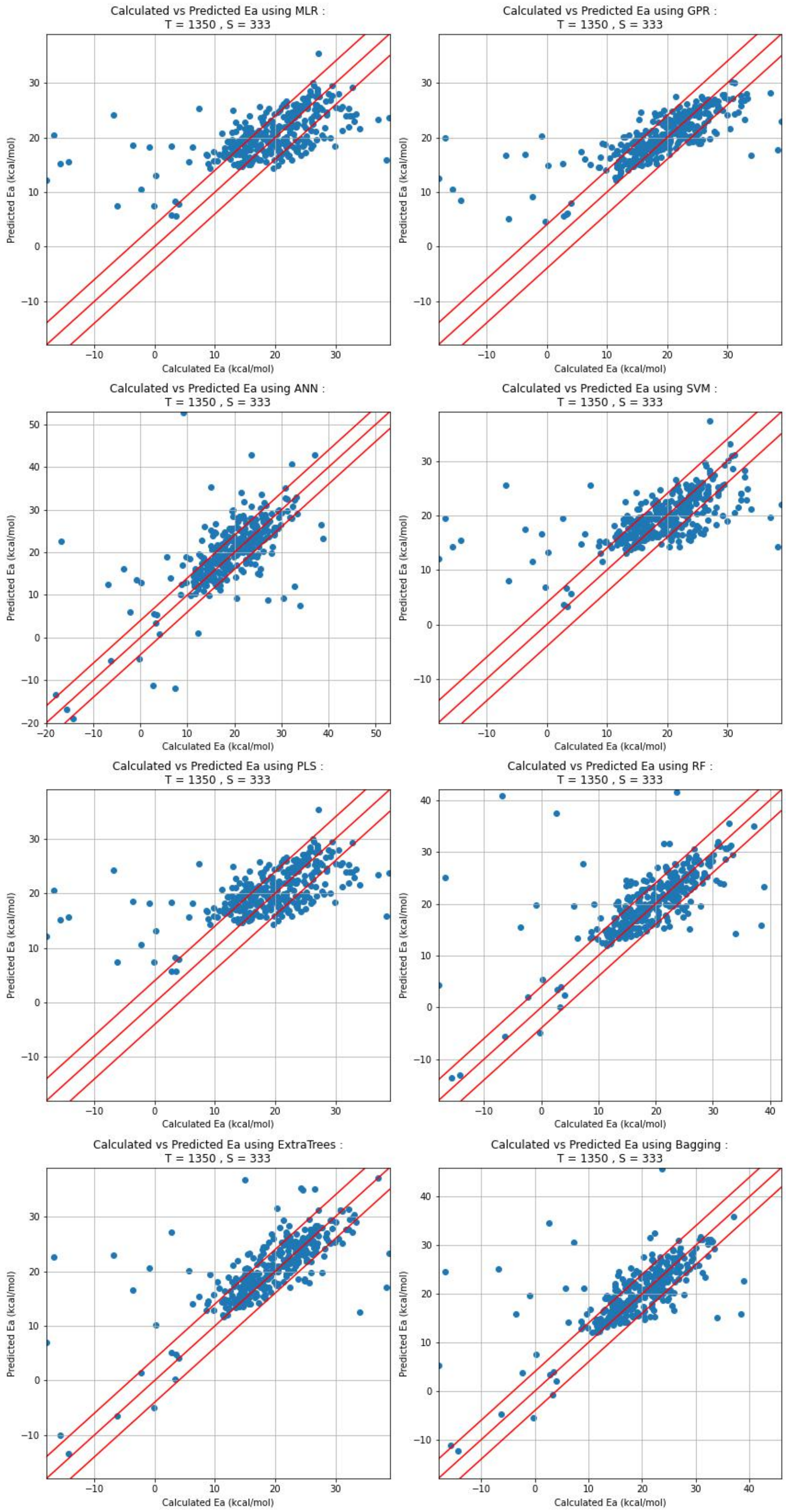
Table 5.A.4: Initial machine learning metrics for the ***ligands_CSD_PIP_set_TSSig*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.28	6.39	67.7	42.1
GPR	0.21	6.89	77.2	55.6
ANN	0.36	6.23	73.2	49.9
SVM	0.31	6.28	80.4	62.2
PLS	0.26	6.47	70.1	44.8
RF	0.39	5.92	77.3	57.3
ExtraTrees	0.39	5.93	77.9	59.9
Bagging	0.39	5.92	77.3	57.3

Table 5.A.5: Initial machine learning metrics for the ***ligands_CSD_PYR_set_TSSig*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.53	6.24	56.1	28.1
GPR	0.35	7.86	62.1	40.0
ANN	0.52	6.51	63.2	38.9
SVM	0.59	5.80	67.3	46.8
PLS	0.50	6.47	55.3	28.4
RF	0.63	5.57	67.3	44.0
ExtraTrees	0.63	5.52	68.5	46.1
Bagging	0.62	5.58	67.3	44.5

5.A.5 Models with Trimmed Descriptors



Calculated vs Predicted Ea using MLR : T = 1350 , S = 333

Calculated vs Predicted Ea using GPR : T = 1350 , S = 333

Calculated vs Predicted Ea using ANN : T = 1350 , S = 333

Calculated vs Predicted Ea using SVM : T = 1350 , S = 333

Calculated vs Predicted Ea using PLS : T = 1350 , S = 333

Calculated vs Predicted Ea using RF : T = 1350 , S = 333

Calculated vs Predicted Ea using ExtraTrees : T = 1350 , S = 333

Calculated vs Predicted Ea using Bagging : T = 1350 , S = 333

Figure 5.A.9: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PIP_set_TSOA*** dataset.

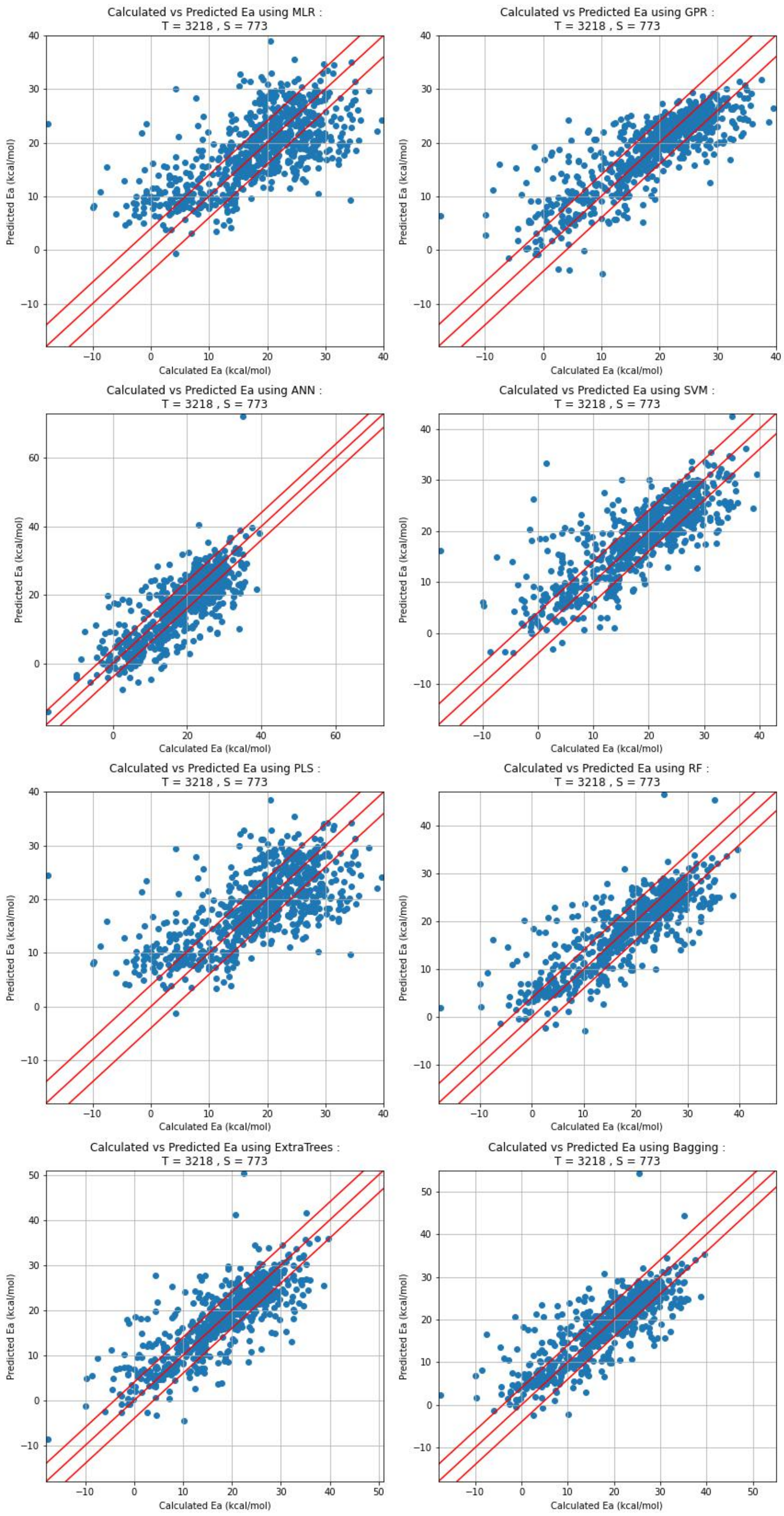


Figure 5.A.10: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSOA*** dataset.

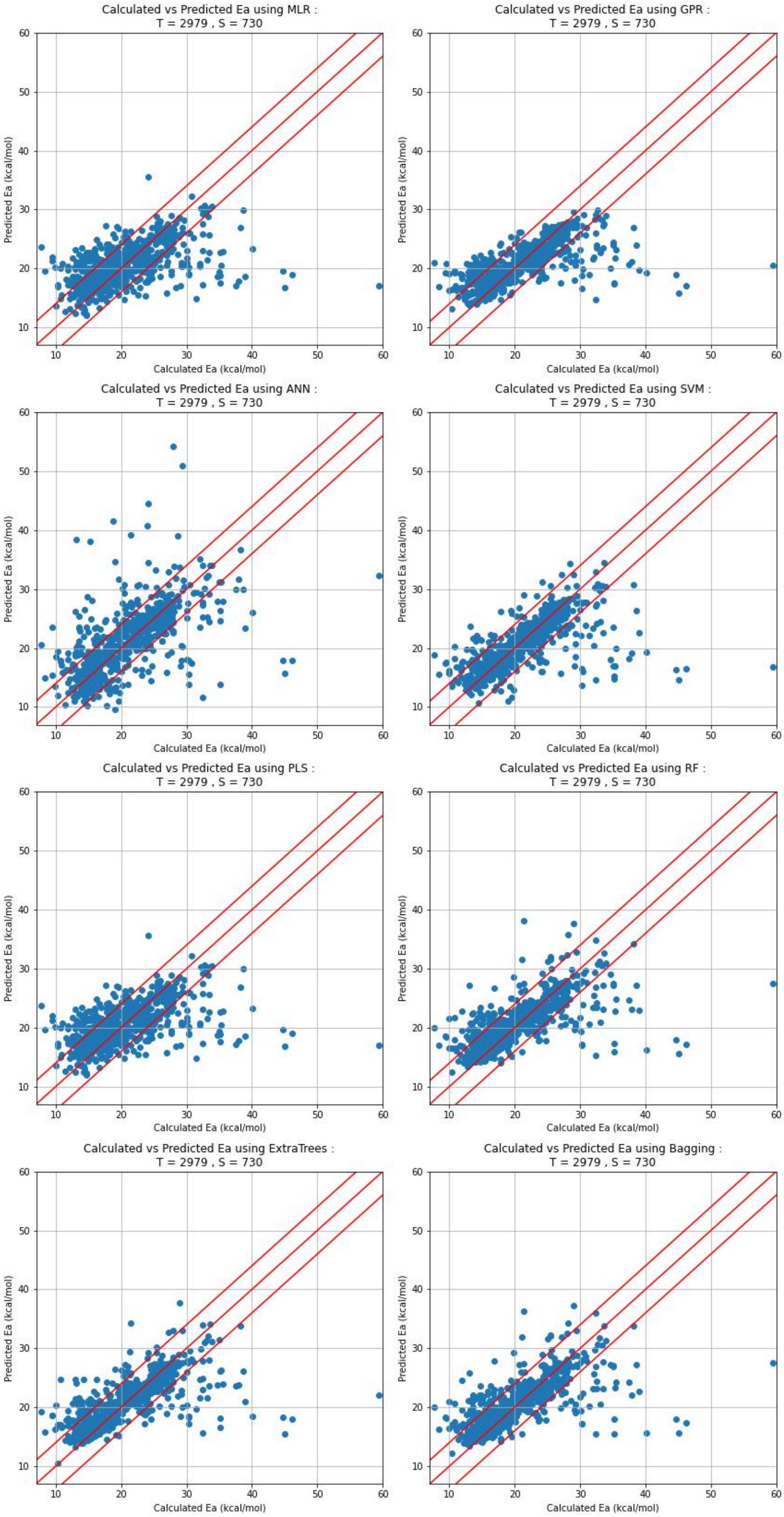


Figure 5.A.11: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PI_set_TSSig*** dataset.

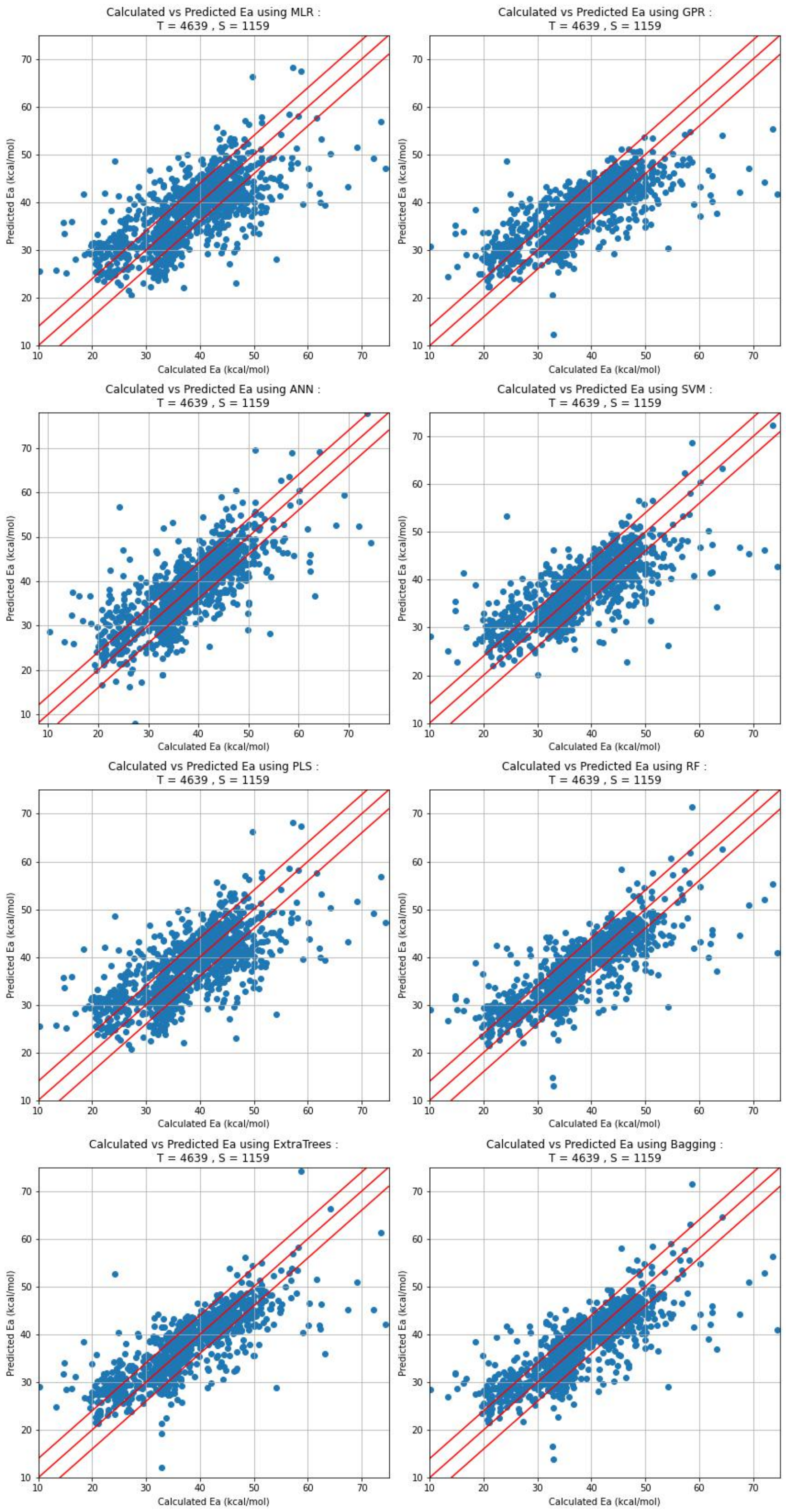


Figure 5.A.12: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSSig*** dataset.

Table 5.A.6: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PIP_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.49	5.85	59.5	30.6
GPR	0.57	5.65	73.9	46.5
ANN	0.54	5.96	66.7	42.6
SVM	0.57	5.28	76.9	49.8
PLS	0.49	5.85	60.4	28.8
RF	0.61	5.09	76.0	55.0
ExtraTrees	0.66	4.78	79.0	57.4
Bagging	0.62	5.02	75.7	55.0

Table 5.A.7: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.44	6.81	50.6	26.8
GPR	0.68	5.17	68.0	46.6
ANN	0.61	5.98	60.4	38.3
SVM	0.64	5.39	66.4	42.3
PLS	0.44	6.81	50.8	28.1
RF	0.69	5.06	70.2	48.6
ExtraTrees	0.67	5.16	70.4	50.5
Bagging	0.68	5.08	69.9	48.3

Table 5.A.8: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PIP_set_TSSig*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.32	4.97	73.7	45.3
GPR	0.45	4.49	79.7	58.6
ANN	0.39	5.09	76.7	53.8
SVM	0.45	4.53	84.7	64.2
PLS	0.32	4.97	73.7	45.3
RF	0.46	4.41	81.0	60.0
ExtraTrees	0.48	4.33	81.1	62.7
Bagging	0.46	4.41	79.7	60.8

Table 5.A.9: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSSig*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.48	6.16	55.9	28.1
GPR	0.60	5.39	66.3	42.5
ANN	0.60	5.54	64.3	41.0
SVM	0.58	5.52	65.1	42.3
PLS	0.48	6.16	55.9	28.1
RF	0.65	5.00	70.3	46.9
ExtraTrees	0.66	4.97	70.9	47.5
Bagging	0.66	4.97	70.1	46.9

5.A.6 Models with Optimised Hyperparameters

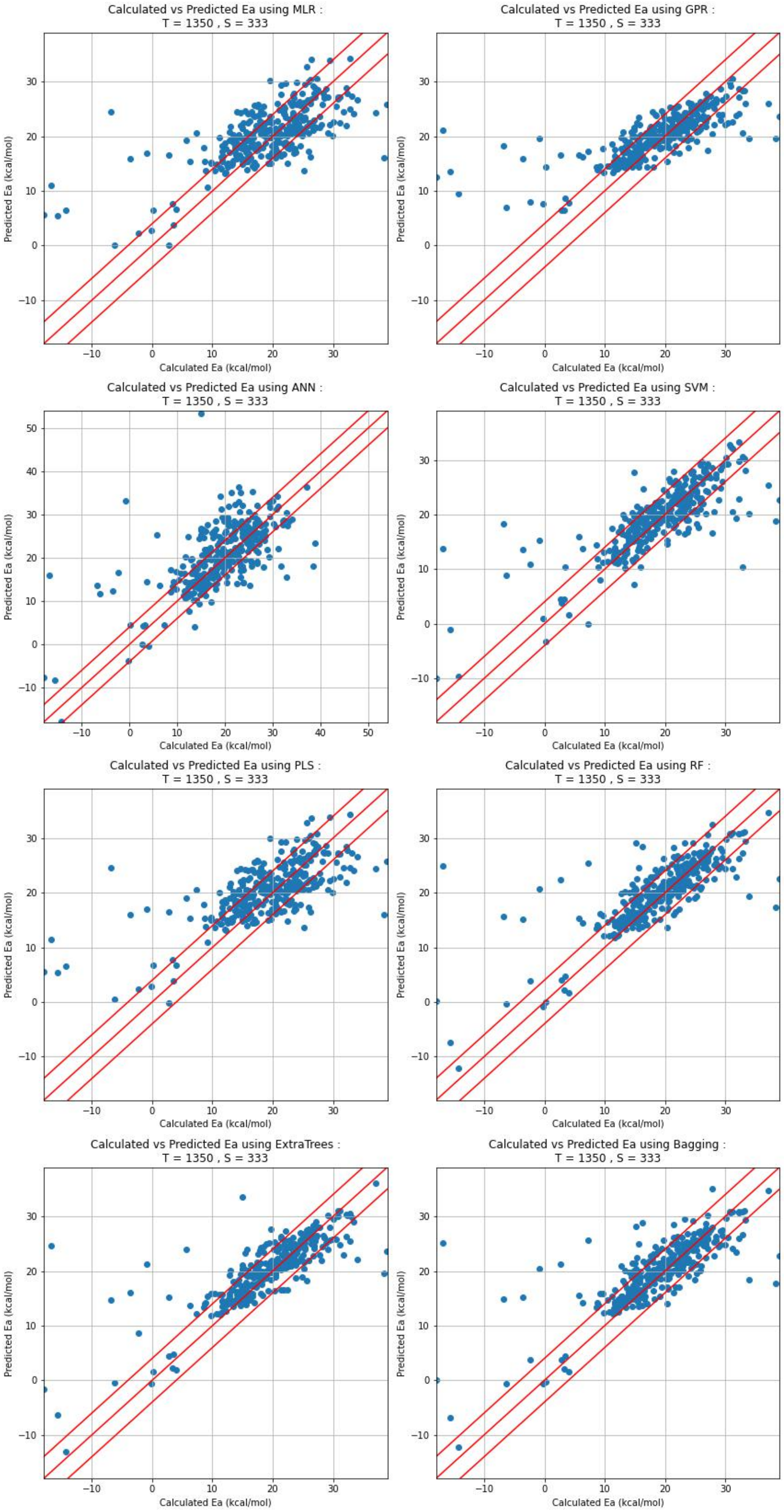


Figure 5.A.13: Machine learning models with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PIP_set_TSOA*** dataset.

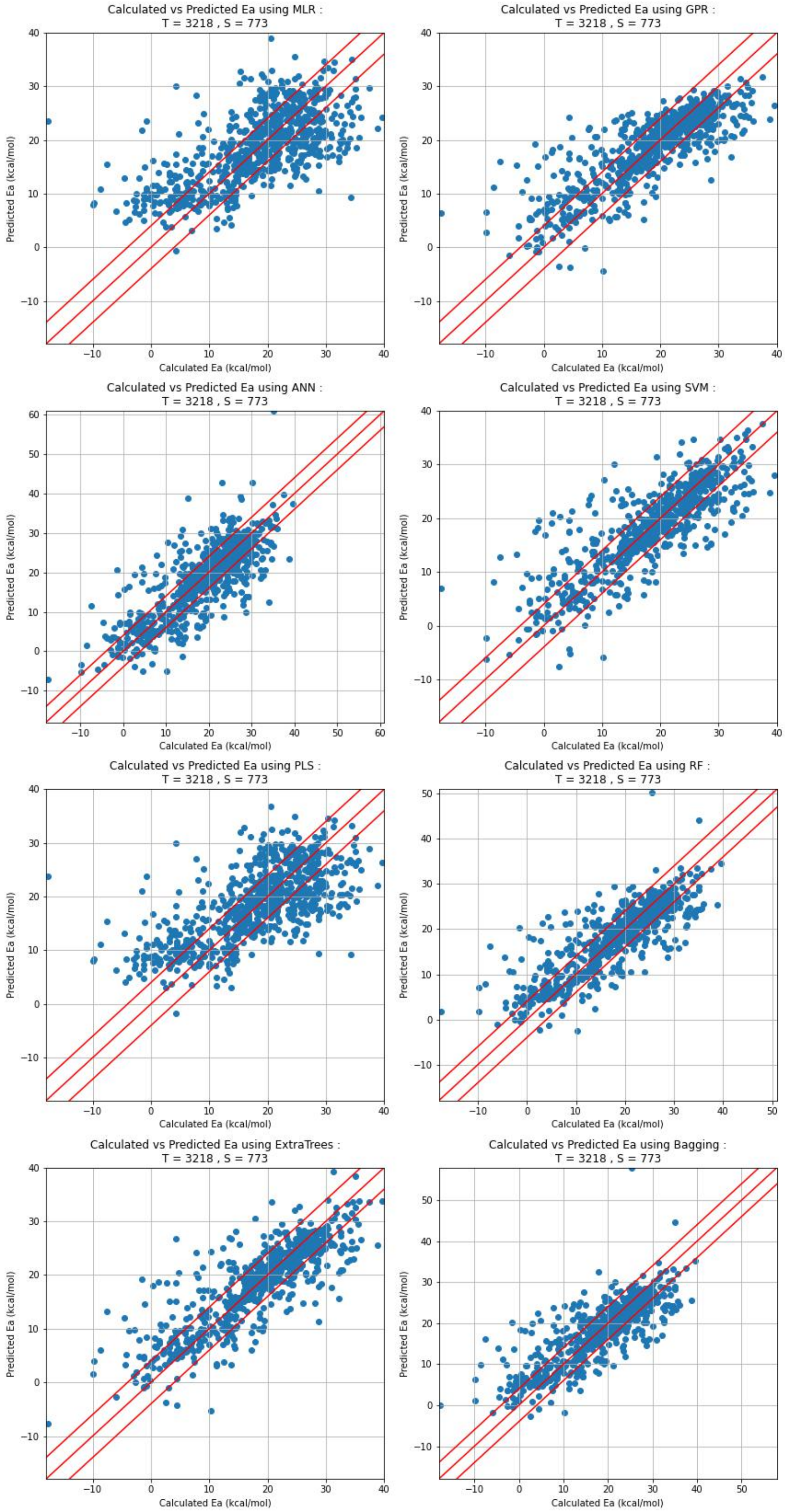


Figure 5.A.14: Machine learning models with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PYR_set_TSOA*** dataset.

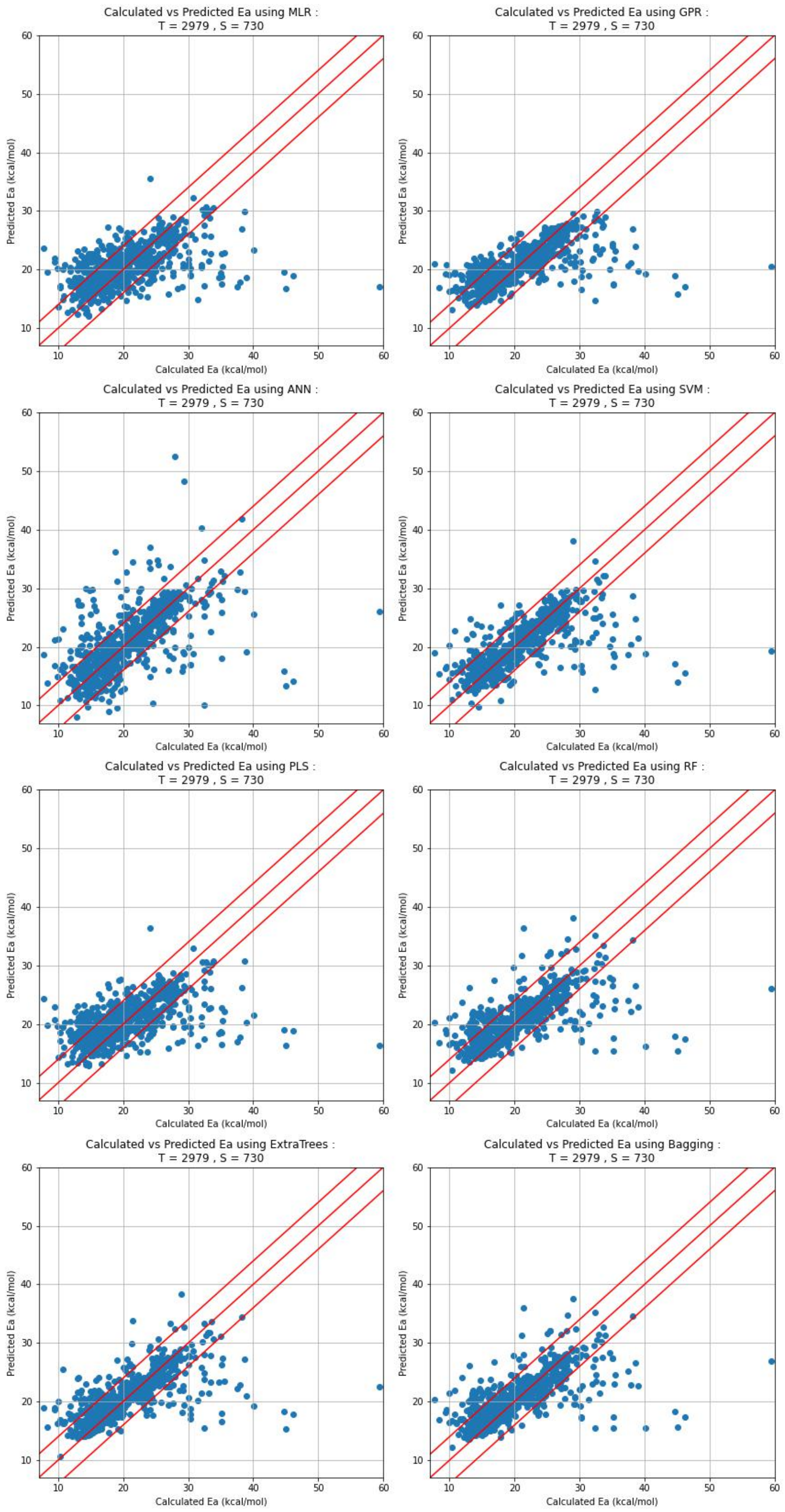


Figure 5.A.15: Machine learning models with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PI_set_TSSig*** dataset.

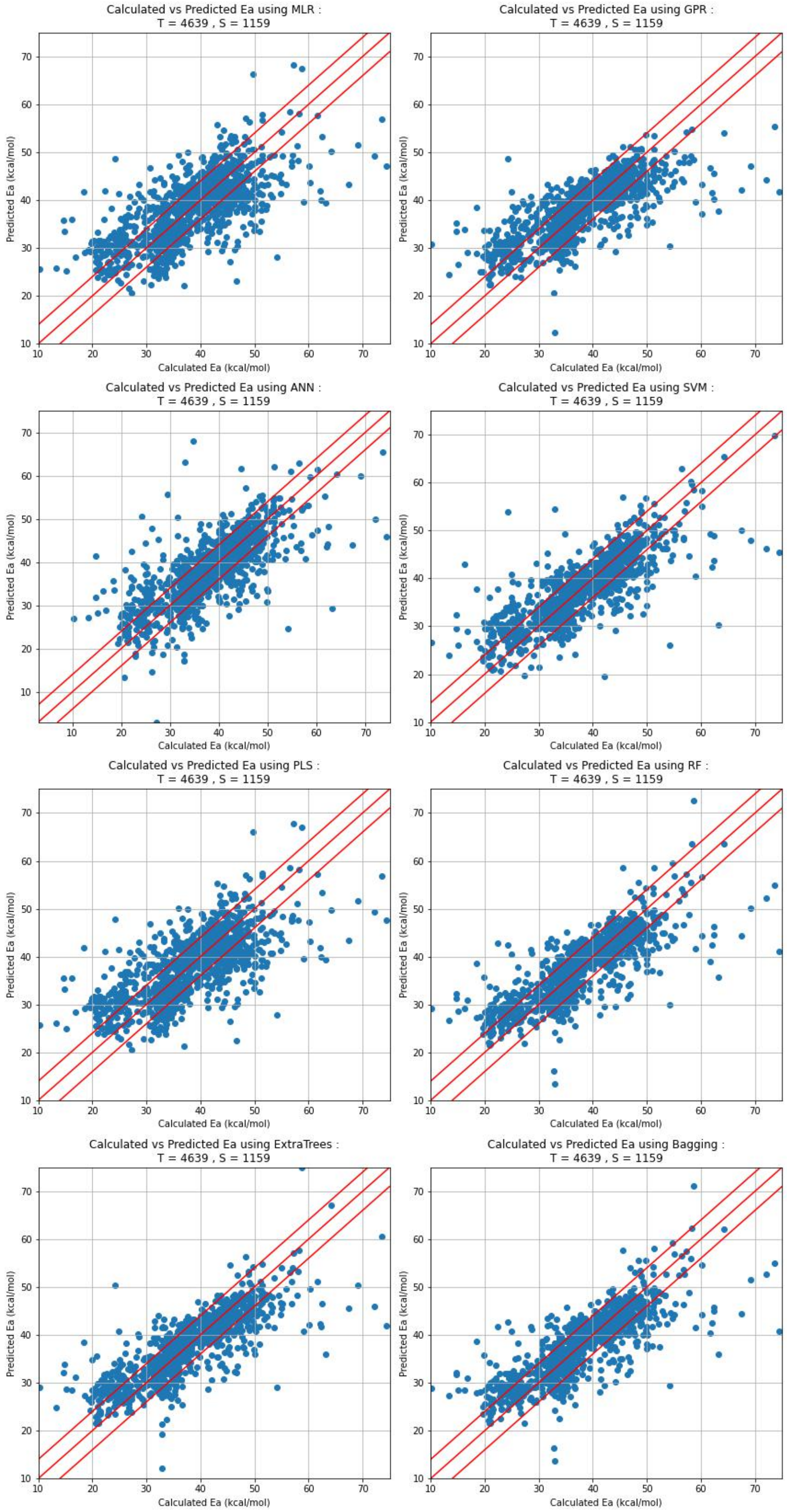


Figure 5.A.16: Machine learning models with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PYR_set_TSSig*** dataset.

Table 5.A.10: Machine learning metrics with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PIP_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.49	5.85	59.5	30.6
GPR	0.57	5.65	73.9	46.5
ANN	0.43	6.54	61.6	37.2
SVM	0.64	4.81	78.7	52.3
PLS	0.49	5.86	60.1	29.7
RF	0.62	5.01	75.7	53.2
ExtraTrees	0.66	4.81	79.6	58.6
Bagging	0.63	5.00	76.6	53.5

Table 5.A.11: Machine learning metrics with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PYR_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.44	6.81	50.6	26.8
GPR	0.68	5.17	68.0	46.6
ANN	0.64	5.63	66.1	41.7
SVM	0.68	5.09	71.7	51.4
PLS	0.44	6.77	51.2	29.0
RF	0.69	5.06	70.2	48.4
ExtraTrees	0.71	4.86	71.3	51.2
Bagging	0.68	5.10	70.2	48.5

Table 5.A.12: Machine learning metrics with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PIP_set_TSSig*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.32	4.97	73.7	45.3
GPR	0.45	4.49	79.7	58.6
ANN	0.41	4.95	78.1	56.8
SVM	0.47	4.41	84.7	66.7
PLS	0.32	4.98	73.8	46.8
RF	0.46	4.44	80.7	61.0
ExtraTrees	0.48	4.33	81.5	62.6
Bagging	0.47	4.39	81.0	61.0

Table 5.A.13: Machine learning metrics with trimmed descriptors and optimised hyperparameters for the ***ligands_CSD_PYR_set_TSSig*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.48	6.16	55.9	28.1
GPR	0.60	5.39	66.3	42.5
ANN	0.55	5.89	65.2	36.9
SVM	0.64	5.10	71.0	48.2
PLS	0.48	6.17	55.8	27.9
RF	0.65	4.99	70.6	47.9
ExtraTrees	0.66	4.95	70.6	47.5
Bagging	0.66	4.97	70.8	47.2

5.A.7 TS Independent Models

5.A.7.1 Models with all descriptors

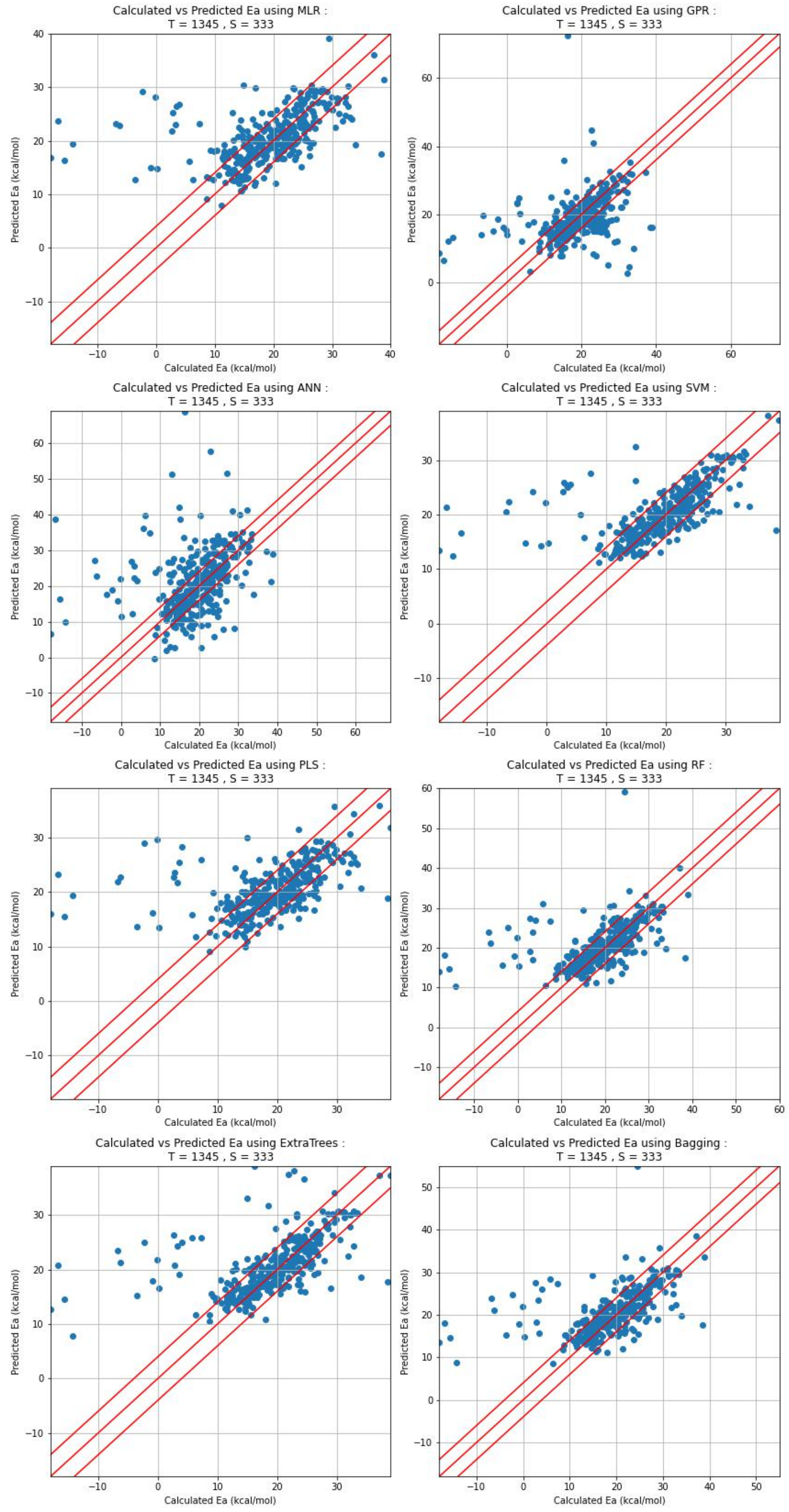


Figure 5.A.17: Initial machine learning models for the ***ligands_CSD_PIP_set_TSOA_NoTS*** dataset.

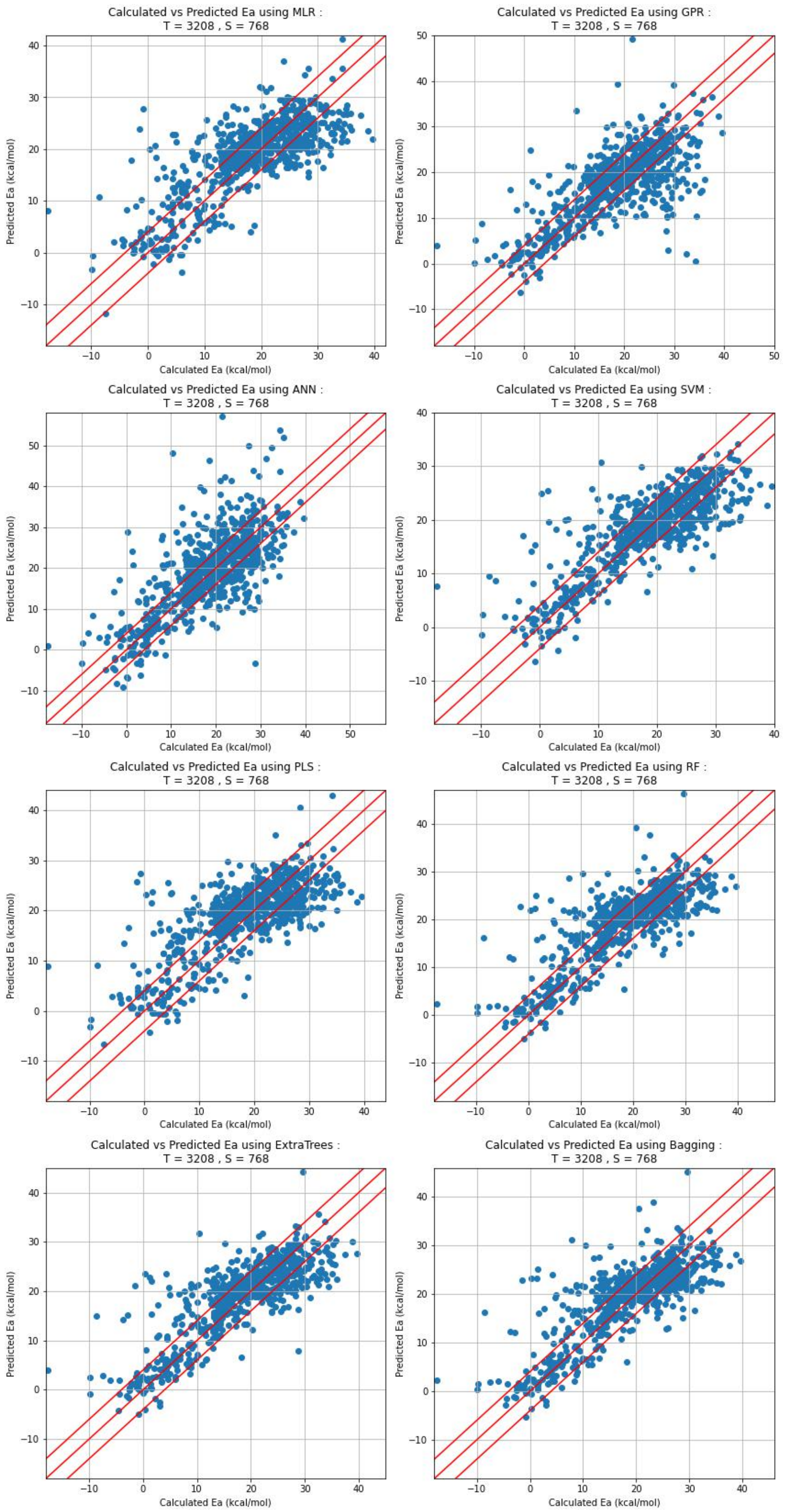


Figure 5.A.18: Initial machine learning models for the ***ligands_CSD_PYR_set_TSOA_-NoTS*** dataset.

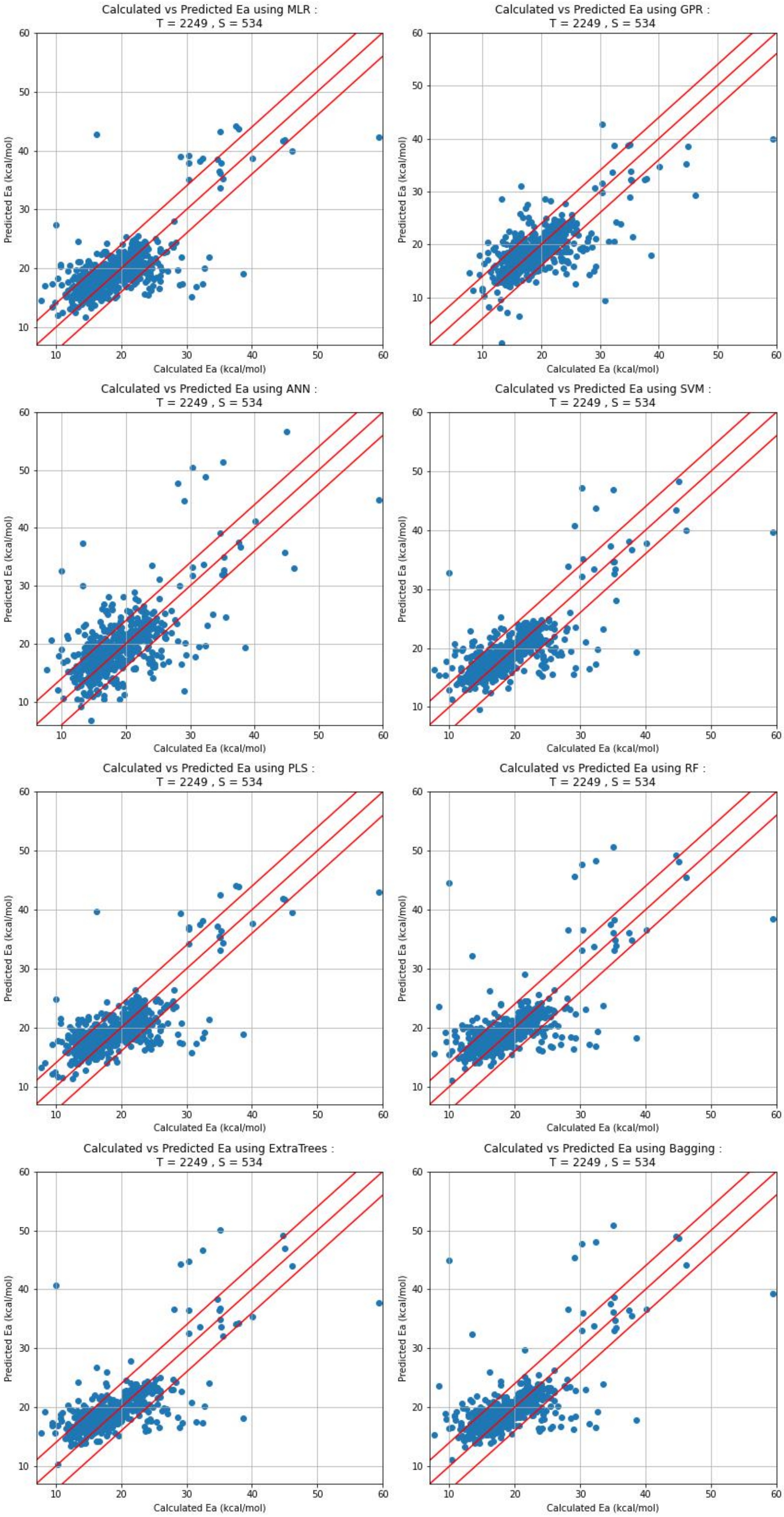


Figure 5.A.19: Initial machine learning models for the ***ligands_CSD_PI_set_TSSig_NoTS*** dataset.

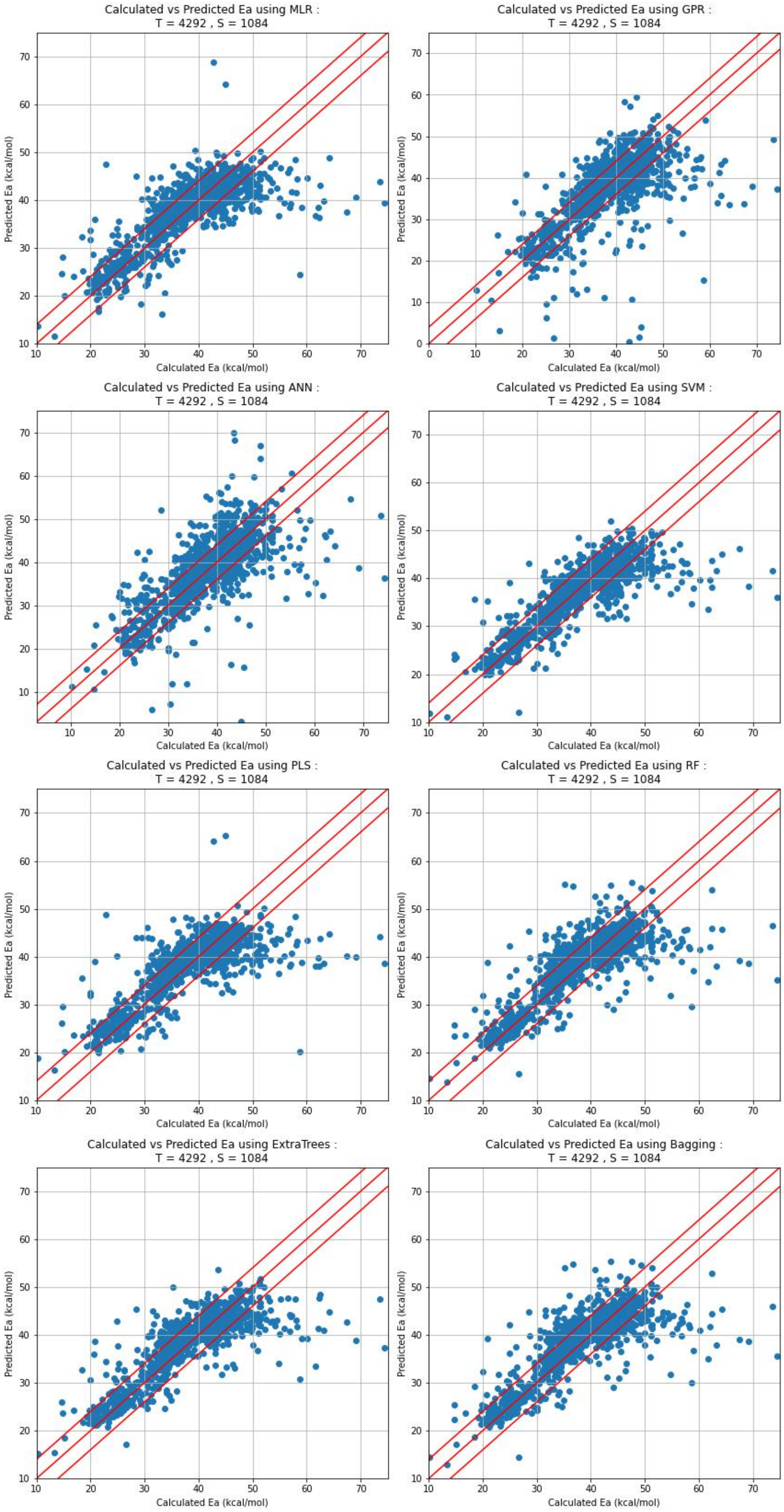


Figure 5.A.20: Initial machine learning models for the ***ligands_CSD_PYR_set_TSSig_-NoTS*** dataset.

Table 5.A.14: Initial machine learning metrics for the ***ligands_CSD_PIP_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.22	7.24	66.7	40.2
GPR	0.15	8.07	64.3	41.1
ANN	0.08	10.59	46.5	26.7
SVM	0.35	6.56	79.6	55.3
PLS	0.23	7.13	65.2	39.0
RF	0.28	7.01	73.0	45.6
ExtraTrees	0.32	6.81	76.0	51.4
Bagging	0.28	6.96	73.6	45.6

Table 5.A.15: Initial machine learning metrics for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.59	5.86	54.8	28.5
GPR	0.55	6.17	62.6	37.4
ANN	0.53	6.98	55.1	32.4
SVM	0.67	5.16	68.1	42.6
PLS	0.56	6.03	54.2	30.5
RF	0.66	5.36	67.1	41.4
ExtraTrees	0.68	5.18	67.7	43.2
Bagging	0.65	5.40	66.8	42.3

Table 5.A.16: Initial machine learning metrics for the ***ligands_CSD_PIP_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.53	3.97	79.8	53.2
GPR	0.48	4.14	78.8	55.4
ANN	0.43	4.74	74.5	47.2
SVM	0.57	3.74	83.7	61.4
PLS	0.54	3.87	79.0	52.4
RF	0.50	4.16	83.3	56.6
ExtraTrees	0.54	3.94	82.6	57.7
Bagging	0.50	4.16	83.1	57.7

Table 5.A.17: Initial machine learning metrics for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.57	5.45	66.5	37.2
GPR	0.47	6.65	64.0	41.3
ANN	0.51	6.31	61.3	36.5
SVM	0.66	4.94	75.6	52.0
PLS	0.56	5.54	68.1	38.5
RF	0.62	5.12	73.2	49.3
ExtraTrees	0.66	4.83	75.8	51.0
Bagging	0.62	5.14	73.4	49.1

5.A.7.2 Models with trimmed descriptors

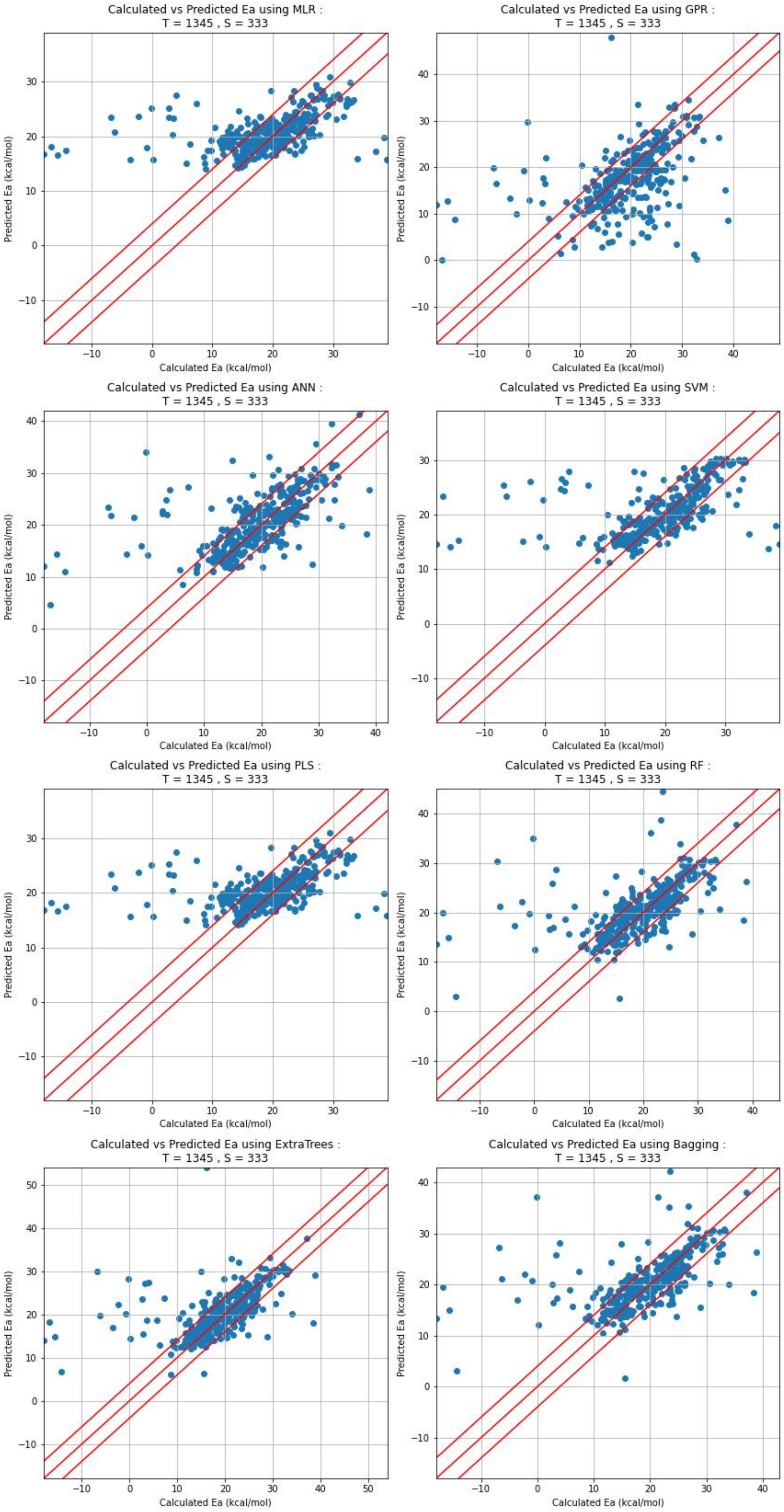


Figure 5.A.21: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PIP_set_TSOA_NoTS*** dataset.

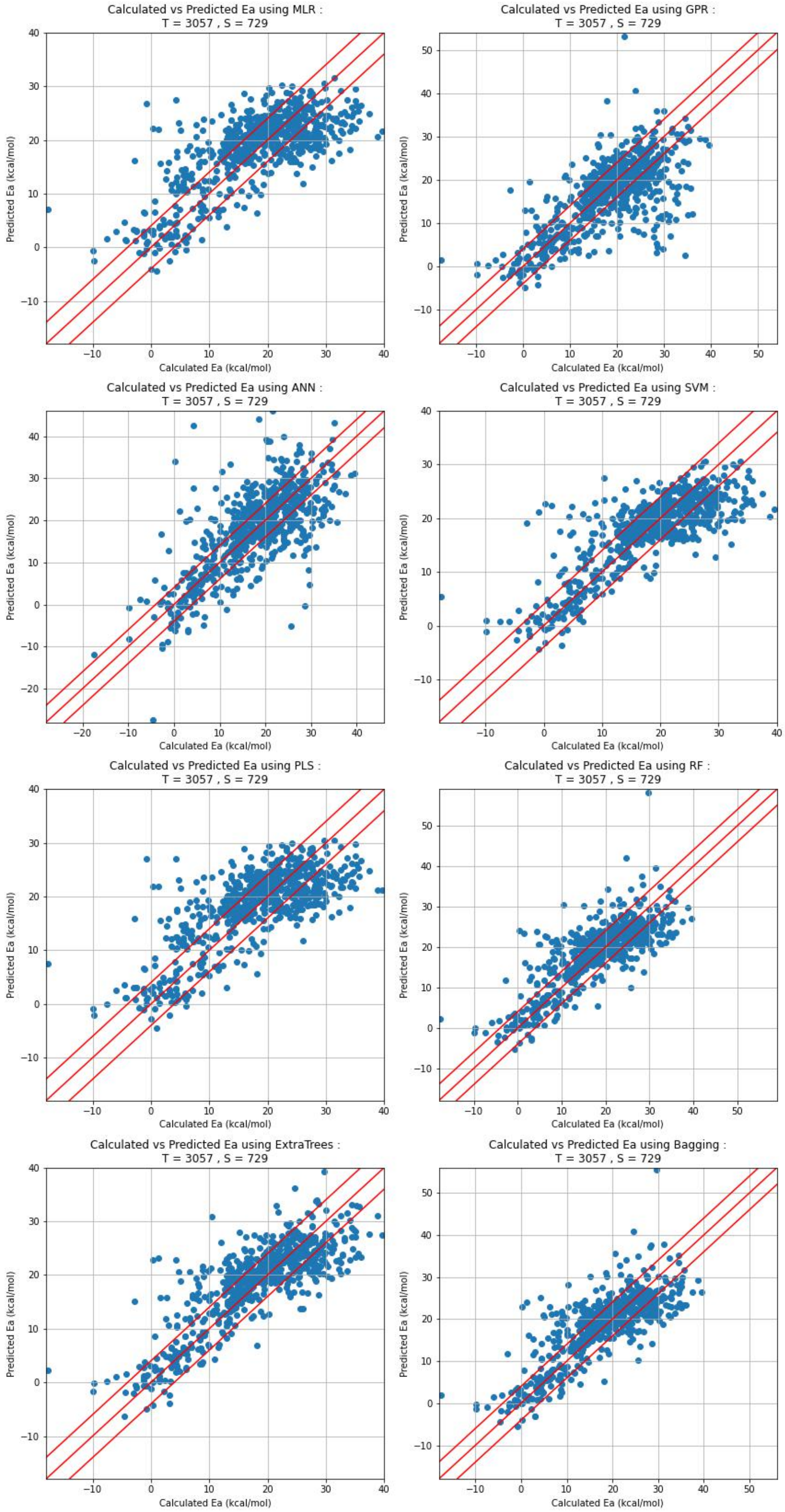


Figure 5.A.22: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSOA_NoTS*** dataset.

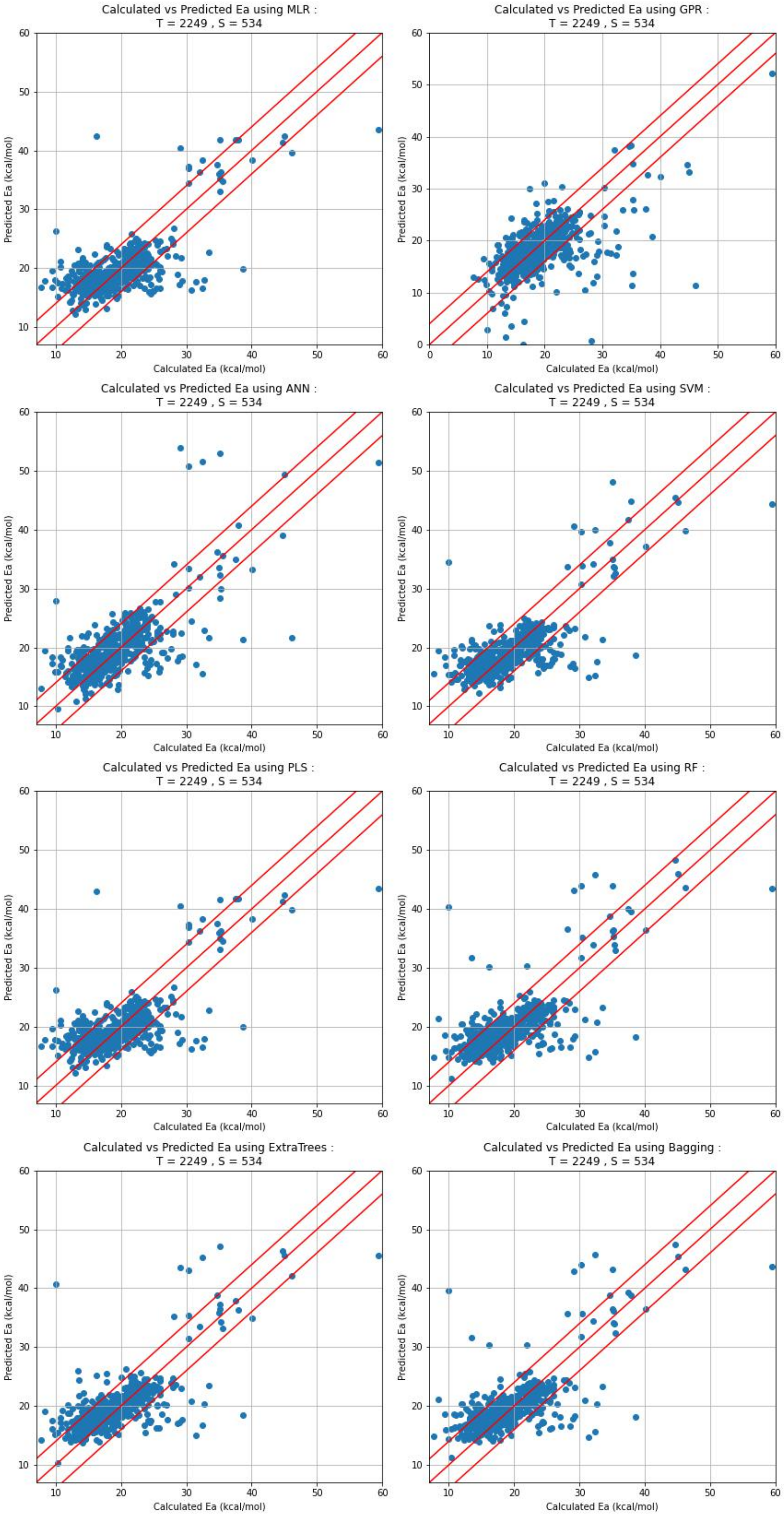


Figure 5.A.23: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PI_set_TSSig_NoTS*** dataset.

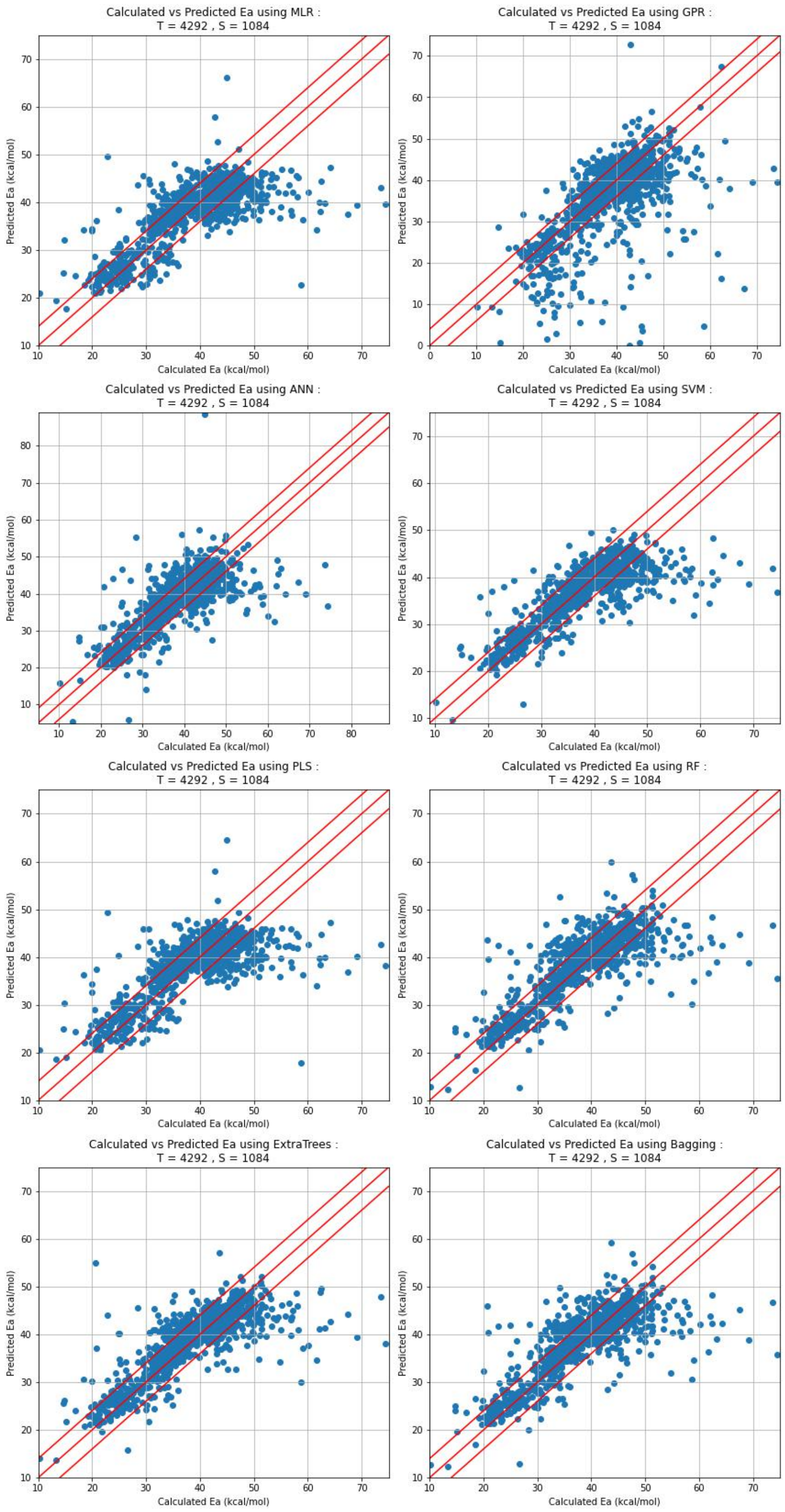


Figure 5.A.24: Machine learning models with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSSig_NoTS*** dataset.

Table 5.A.18: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PIP_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.22	7.22	65.8	40.2
GPR	0.19	8.03	58.3	39.3
ANN	0.33	6.73	68.5	41.1
SVM	0.24	7.07	75.1	47.7
PLS	0.22	7.22	65.8	40.2
RF	0.28	6.99	71.2	46.2
ExtraTrees	0.29	6.95	76.3	49.2
Bagging	0.28	6.95	72.1	48.0

Table 5.A.19: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.55	5.98	52.3	30.0
GPR	0.52	6.46	58.2	36.5
ANN	0.52	6.80	55.1	32.5
SVM	0.65	5.23	66.3	42.1
PLS	0.55	5.95	52.1	29.6
RF	0.66	5.24	66.0	39.9
ExtraTrees	0.69	4.97	66.3	41.2
Bagging	0.66	5.25	65.3	41.3

Table 5.A.20: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PIP_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.50	4.04	77.3	50.7
GPR	0.35	4.84	74.7	49.3
ANN	0.51	4.18	80.1	48.7
SVM	0.56	3.78	82.8	58.2
PLS	0.50	4.04	77.2	51.3
RF	0.53	4.00	82.0	57.7
ExtraTrees	0.56	3.87	83.0	56.6
Bagging	0.53	3.98	82.4	56.9

Table 5.A.21: Machine learning metrics with trimmed descriptors for the ***ligands_CSD_-PYR_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.55	5.56	65.7	37.5
GPR	0.40	7.92	62.2	43.0
ANN	0.57	5.59	71.2	45.1
SVM	0.62	5.21	73.5	50.6
PLS	0.54	5.65	66.2	36.3
RF	0.62	5.16	74.4	51.4
ExtraTrees	0.64	4.98	75.8	49.8
Bagging	0.62	5.16	73.8	50.8

5.A.8 DFT Models

5.A.8.1 TS Independent Models

5.A.8.1.1 PBE0

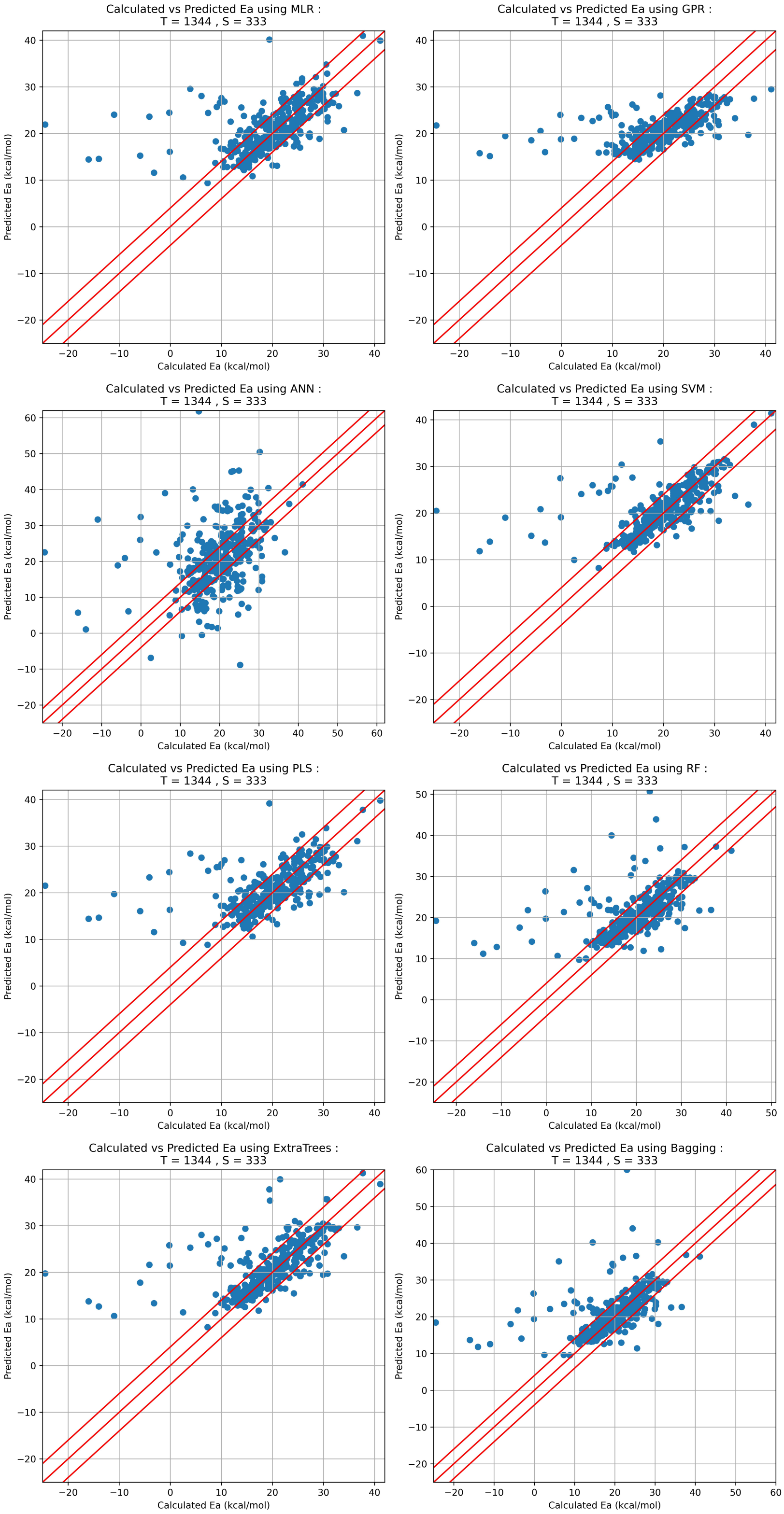
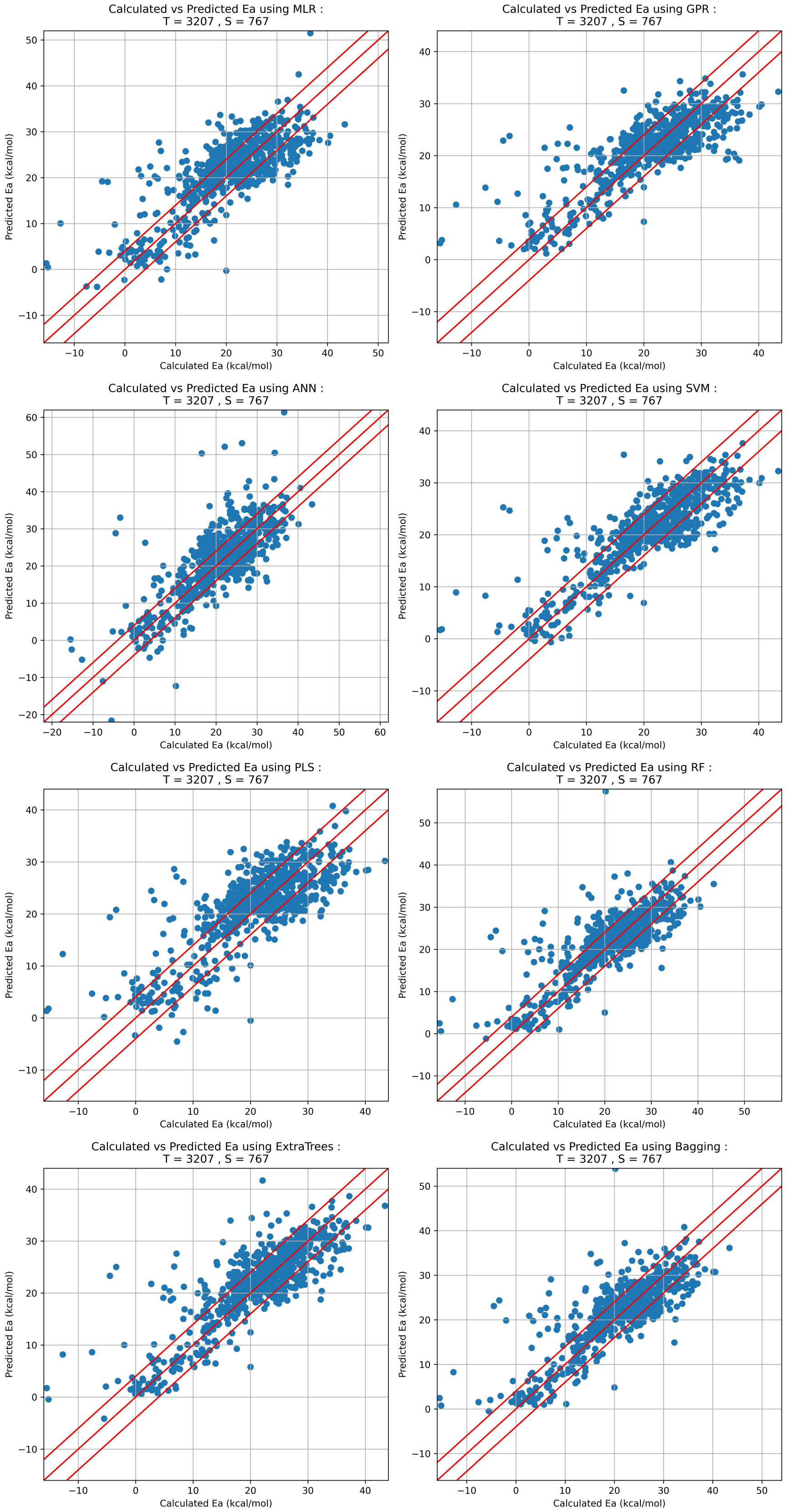


Figure 5.A.25: Machine learning models using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PIP_set_TSOA_NoTS*** dataset.



Calculated vs Predicted Ea using MLR : T = 3207 , S = 767

Calculated vs Predicted Ea using GPR : T = 3207 , S = 767

Calculated vs Predicted Ea using ANN : T = 3207 , S = 767

Calculated vs Predicted Ea using SVM : T = 3207 , S = 767

Calculated vs Predicted Ea using PLS : T = 3207 , S = 767

Calculated vs Predicted Ea using RF : T = 3207 , S = 767

Calculated vs Predicted Ea using ExtraTrees : T = 3207 , S = 767

Calculated vs Predicted Ea using Bagging : T = 3207 , S = 767

Figure 5.A.26: Machine learning models using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

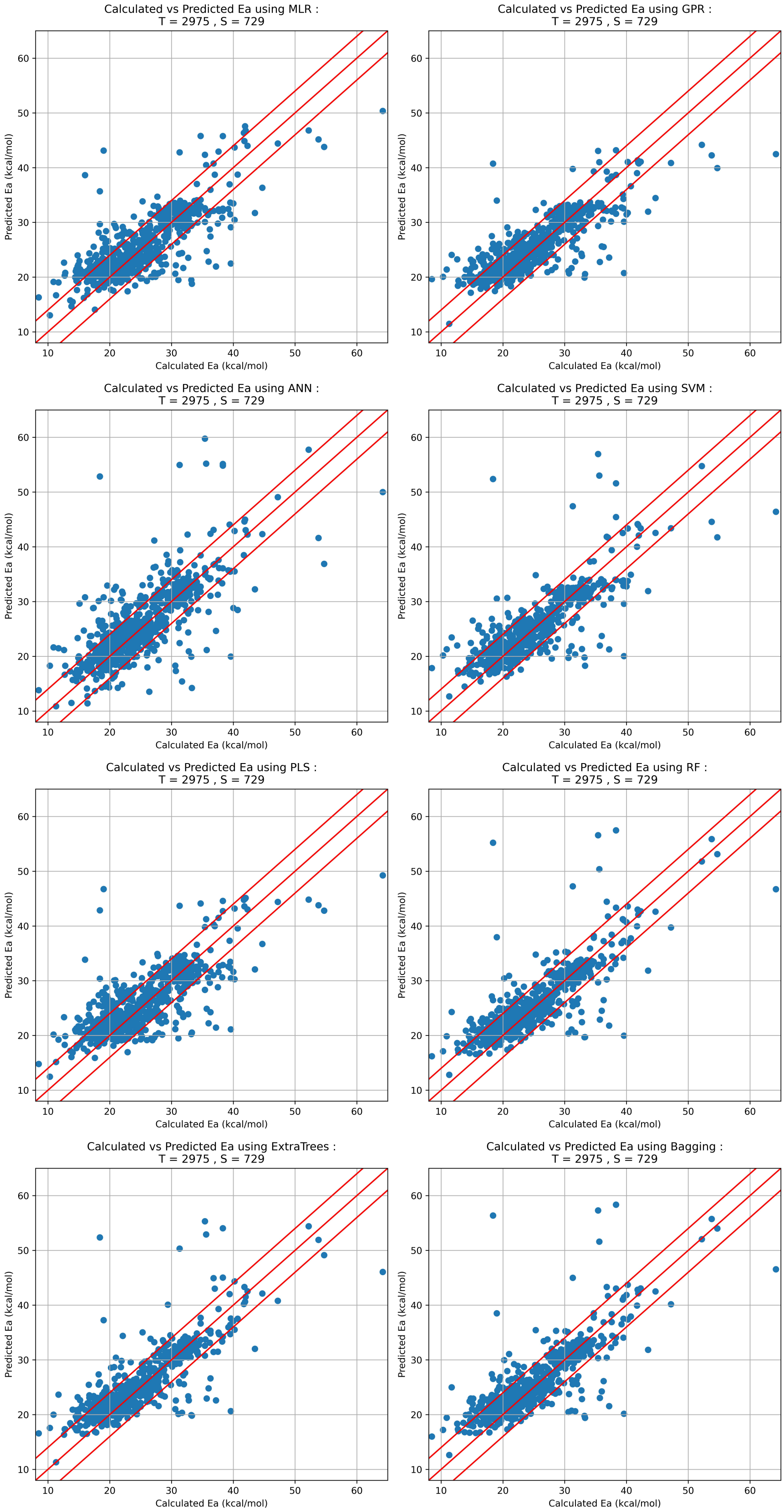


Figure 5.A.27: Machine learning models using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PI_set_TSSig_NoTS*** dataset.

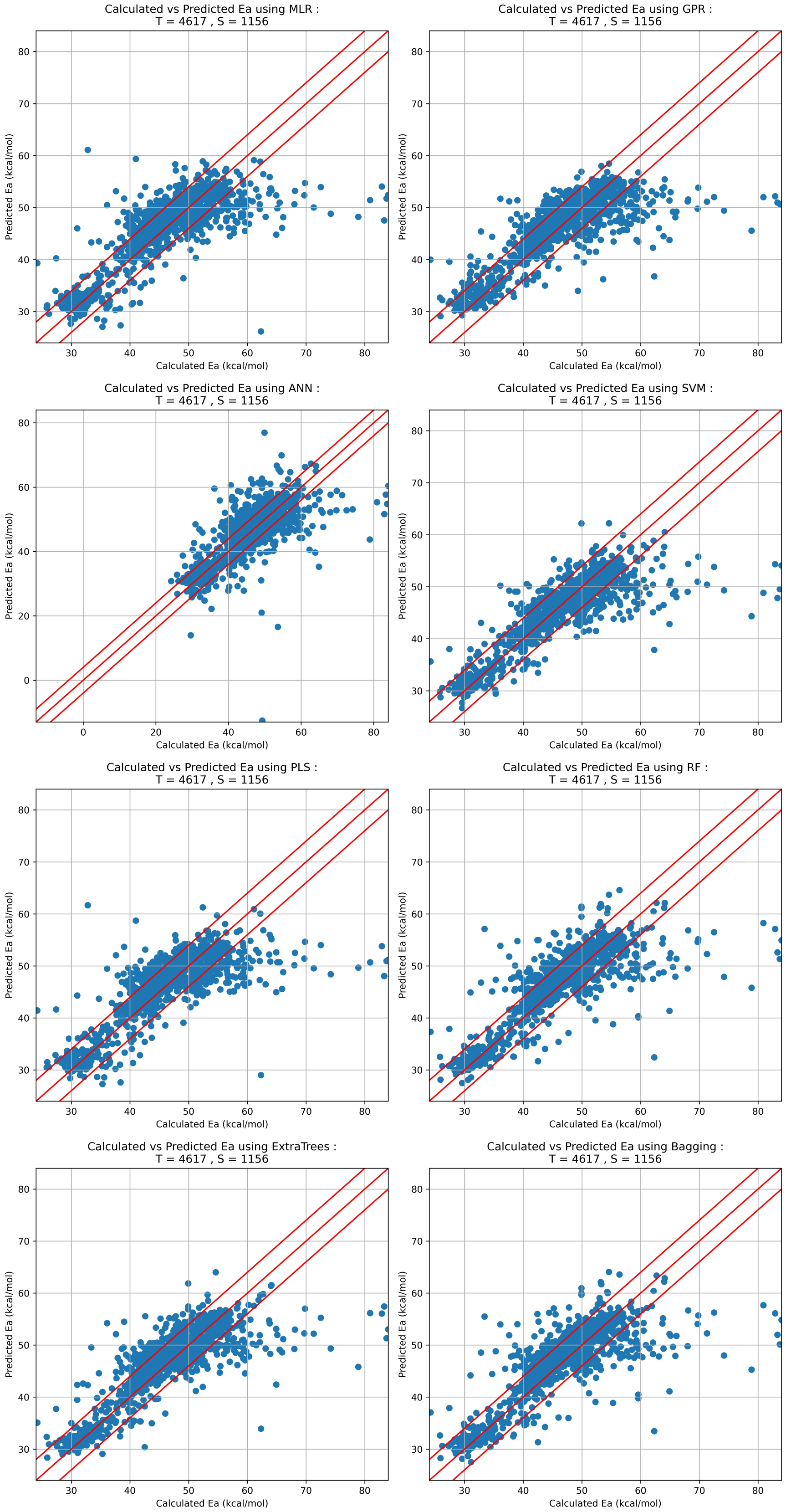


Figure 5.A.28: Machine learning models using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

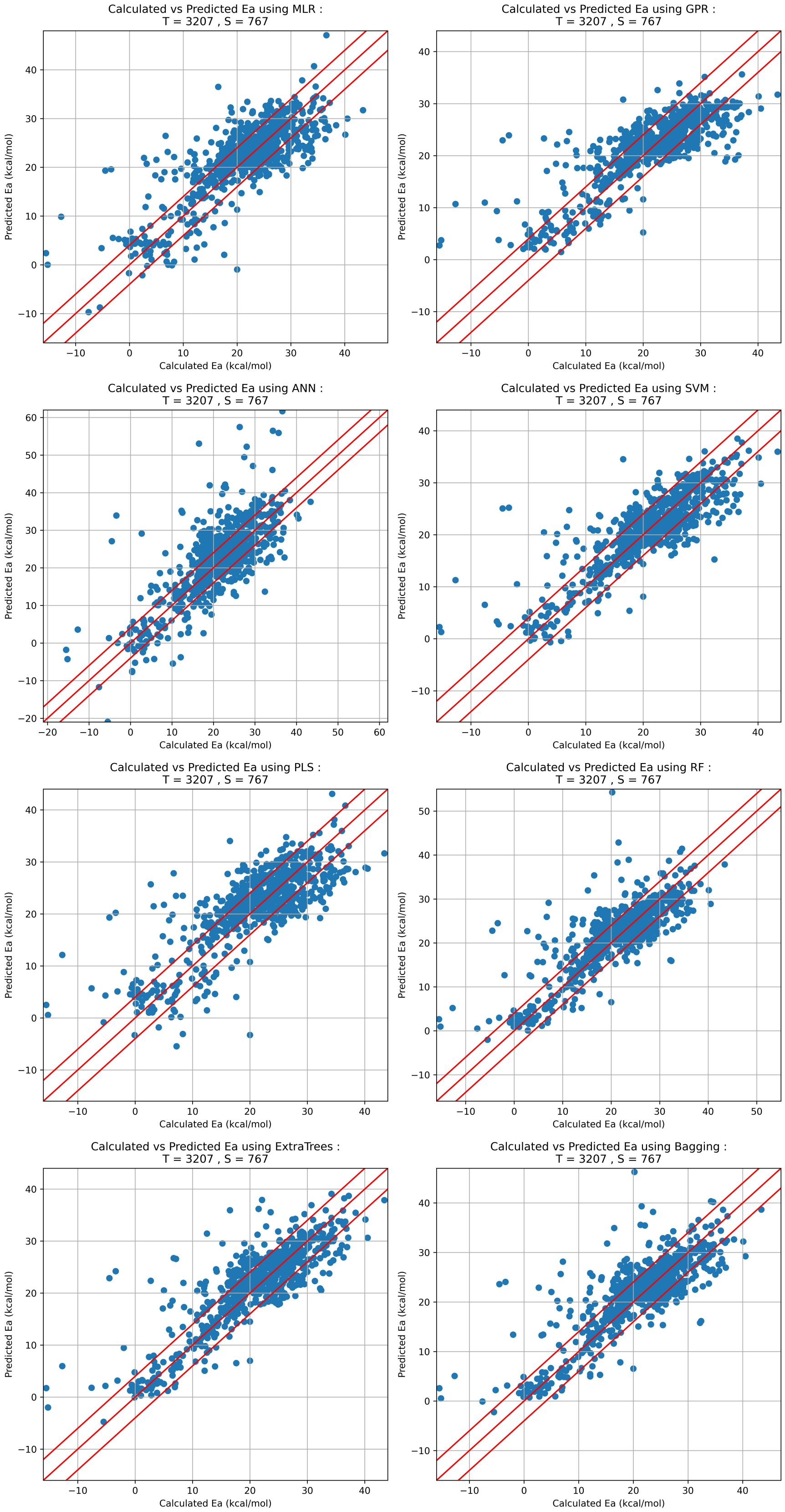


Figure 5.A.29: Machine learning models using an expanded descriptor set calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

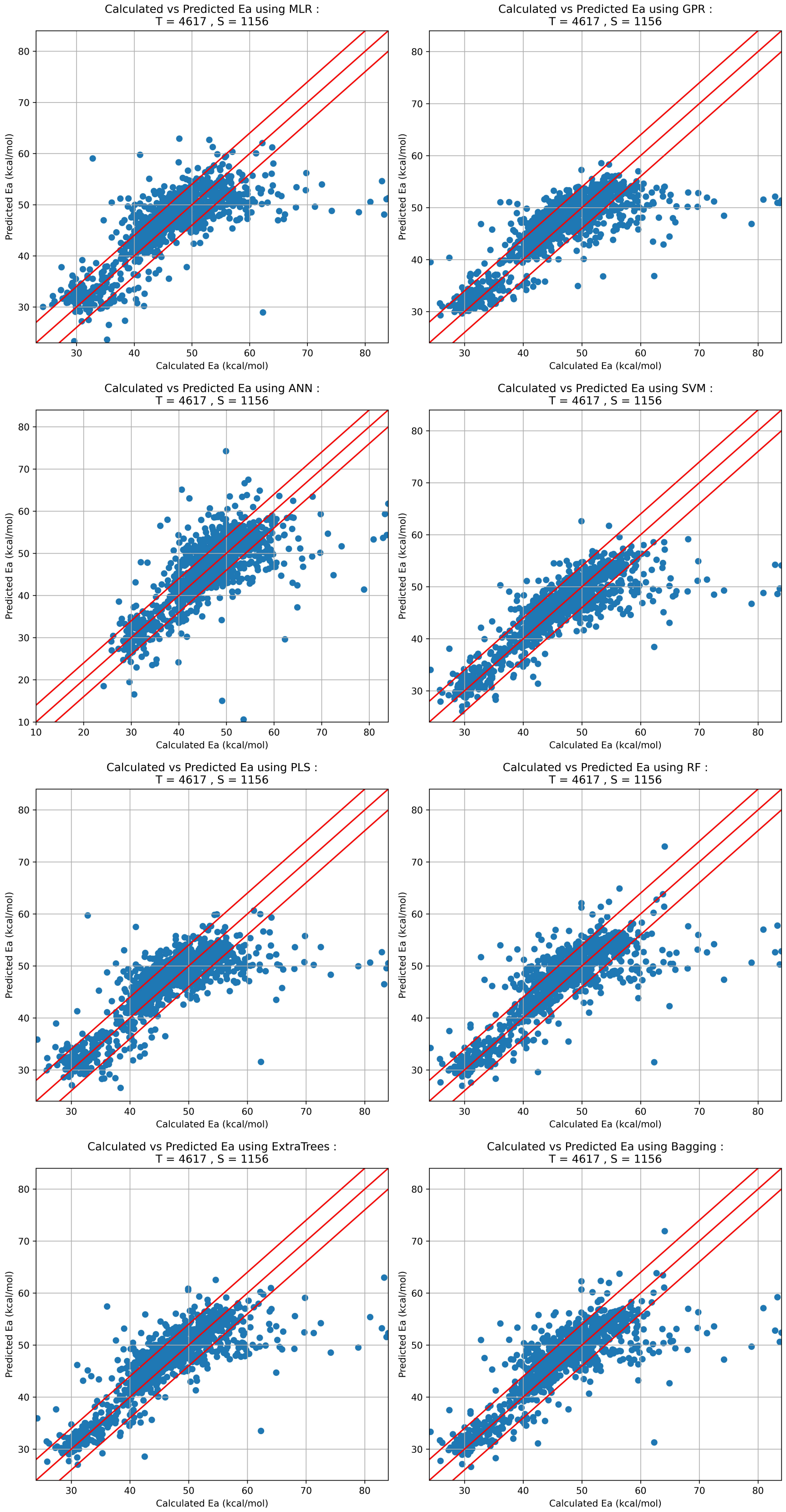


Figure 5.A.30: Machine learning models using and expanded descriptor set calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

Table 5.A.22: Machine learning metrics using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PIP_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.30	6.63	71.8	40.8
GPR	0.34	6.48	71.8	45.6
ANN	0.13	9.62	47.4	21.9
SVM	0.40	6.11	82.3	58.6
PLS	0.32	6.51	73.0	42.9
RF	0.32	6.62	77.2	54.7
ExtraTrees	0.38	6.28	77.8	55.0
Bagging	0.30	6.84	77.5	54.7

Table 5.A.23: Machine learning metrics using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.62	5.35	59.6	29.5
GPR	0.66	5.03	67.3	40.4
ANN	0.62	5.90	62.5	34.3
SVM	0.71	4.59	72.6	47.3
PLS	0.61	5.43	57.2	32.7
RF	0.67	5.00	71.7	46.3
ExtraTrees	0.71	4.68	72.9	46.3
Bagging	0.68	4.97	71.7	45.6

Table 5.A.24: Machine learning metrics using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PIP_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.65	3.99	76.1	50.5
GPR	0.71	3.66	81.8	59.5
ANN	0.60	4.58	76.0	51.9
SVM	0.69	3.78	84.6	62.0
PLS	0.63	4.12	77.2	49.9
RF	0.69	3.80	84.4	61.3
ExtraTrees	0.69	3.79	84.4	63.0
Bagging	0.69	3.83	84.4	61.0

Table 5.A.25: Machine learning metrics using descriptors calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.60	5.16	72.5	41.6
GPR	0.62	5.03	74.7	46.9
ANN	0.51	6.11	66.2	39.4
SVM	0.68	4.72	79.3	56.5
PLS	0.61	5.12	73.8	42.1
RF	0.65	4.85	77.9	55.3
ExtraTrees	0.68	4.66	78.0	54.7
Bagging	0.65	4.88	77.3	54.9

Table 5.A.26: Machine learning metrics using an expanded descriptor set calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.64	5.25	61.4	31.6
GPR	0.67	4.97	66.6	38.5
ANN	0.60	6.44	58.3	32.5
SVM	0.72	4.53	72.9	49.2
PLS	0.62	5.36	58.3	29.7
RF	0.68	4.94	72.5	47.2
ExtraTrees	0.73	4.56	73.5	48.4
Bagging	0.69	4.85	72.2	47.6

Table 5.A.27: Machine learning metrics using an expanded descriptor set calculated using PBE0/def2-TZVP for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.62	5.04	74.5	40.1
GPR	0.64	4.90	76.1	47.8
ANN	0.56	5.81	68.7	42.1
SVM	0.70	4.58	80.7	58.7
PLS	0.62	5.03	73.5	42.9
RF	0.67	4.70	78.4	56.6
ExtraTrees	0.70	4.51	79.1	57.4
Bagging	0.68	4.66	78.7	56.7

5.A.8.1.2 TPSS

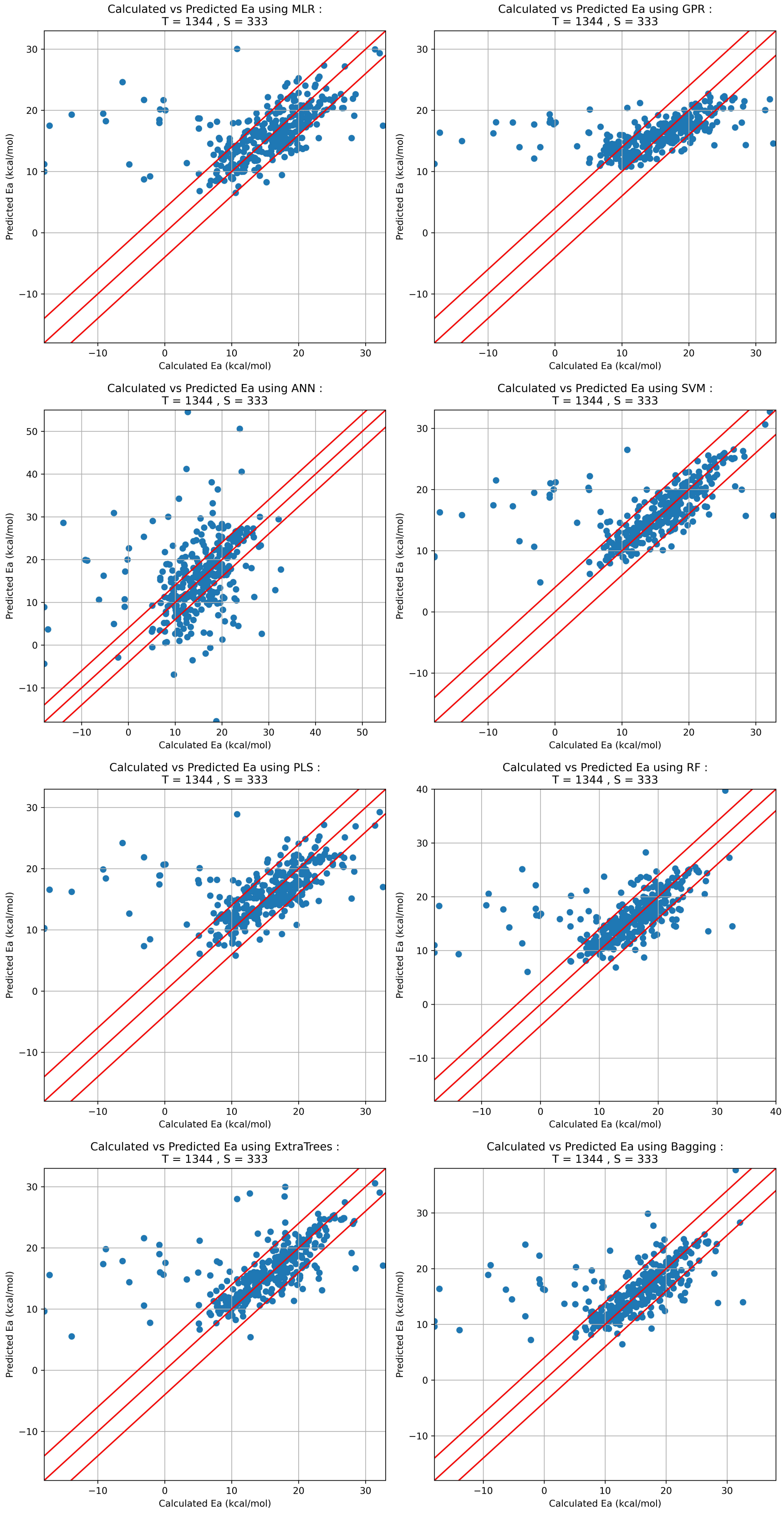


Figure 5.A.31: Machine learning models using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PIP_set_TSOA_NoTS*** dataset.

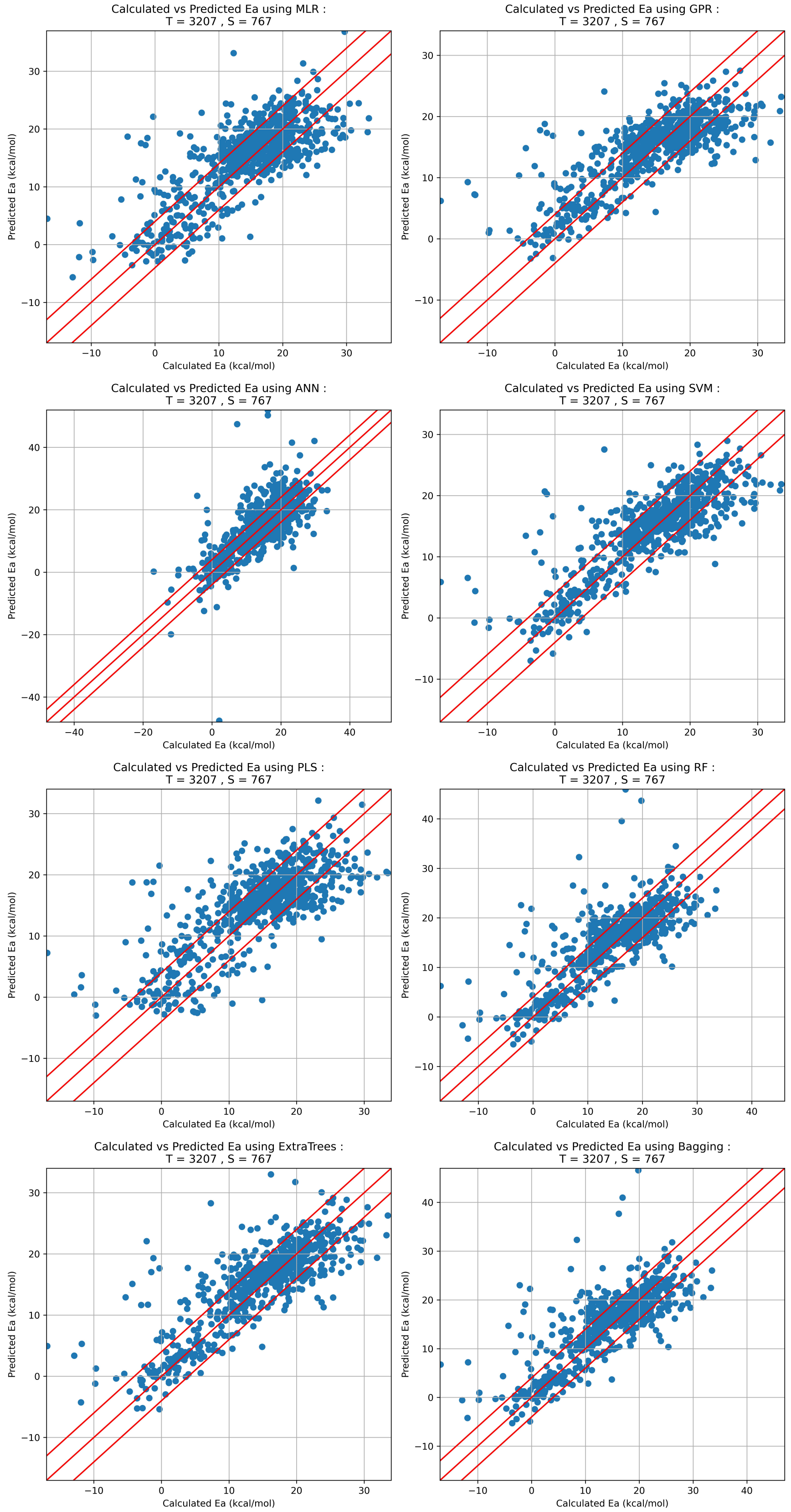


Figure 5.A.32: Machine learning models using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

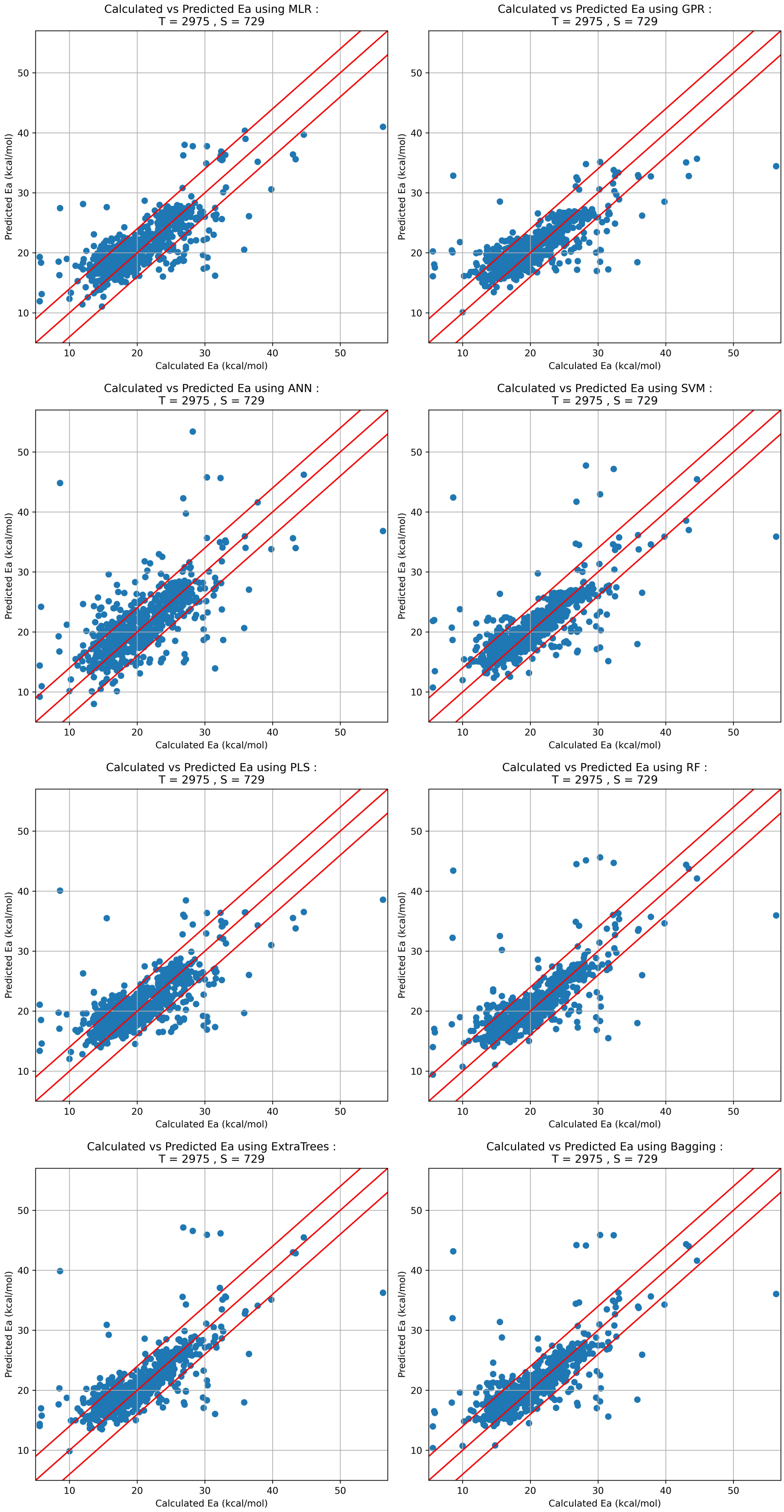


Figure 5.A.33: Machine learning models using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PI_set_TSSig_NoTS*** dataset.

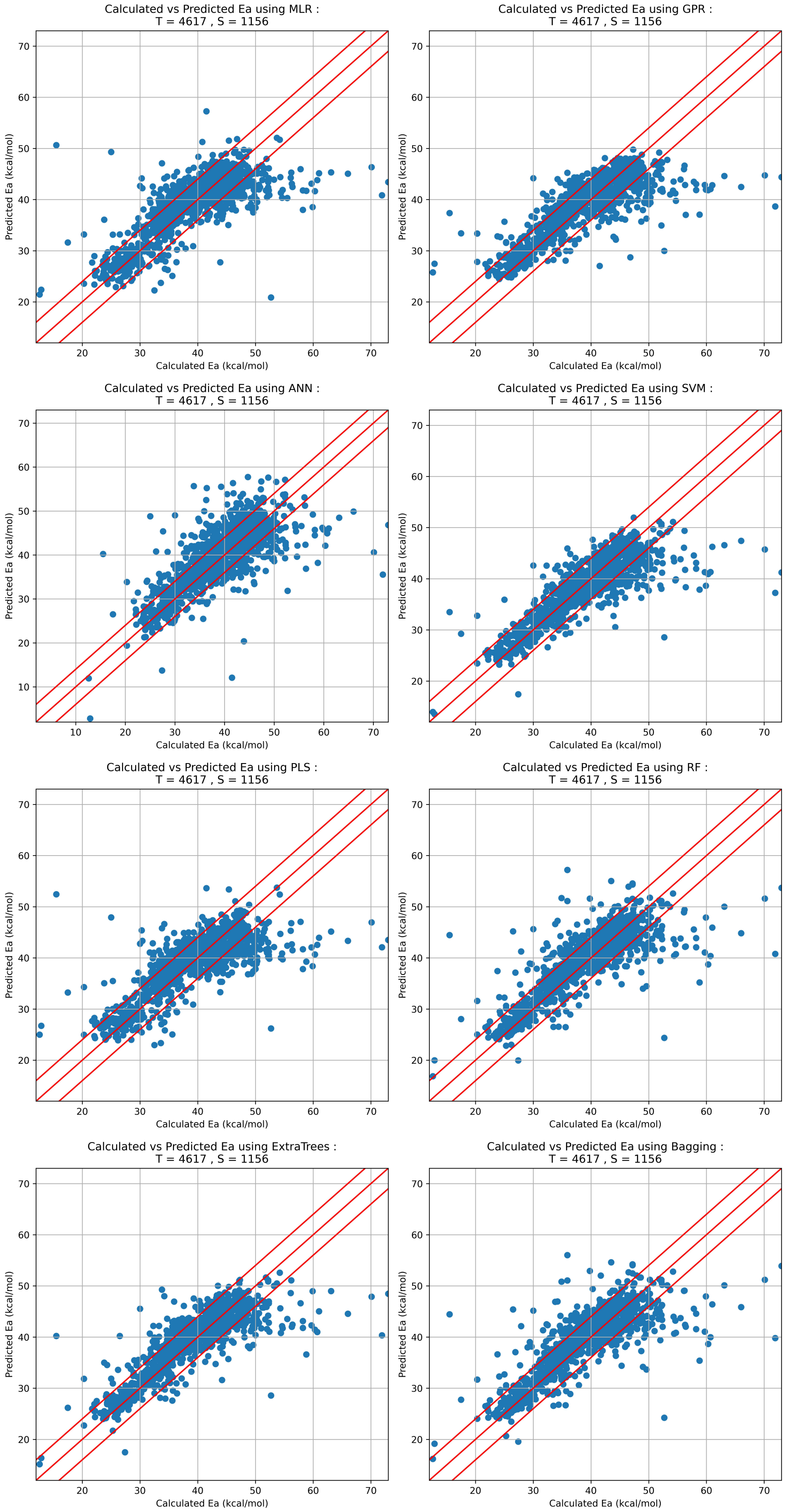


Figure 5.A.34: Machine learning models using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

Table 5.A.28: Machine learning metrics using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PIP_set_TSOA*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.22	6.54	74.5	40.8
GPR	0.27	6.36	70.9	45.3
ANN	0.10	9.31	51.7	33.3
SVM	0.33	6.07	82.3	62.2
PLS	0.24	6.44	73.9	44.4
RF	0.29	6.26	77.8	55.6
ExtraTrees	0.34	6.03	80.5	58.0
Bagging	0.29	6.23	77.5	56.8

Table 5.A.29: Machine learning metrics using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PYR_set_TSOA_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.59	5.01	64.1	34.4
GPR	0.64	4.69	69.6	45.1
ANN	0.53	6.05	63.0	39.1
SVM	0.69	4.27	75.4	52.3
PLS	0.58	5.09	62.3	34.6
RF	0.63	4.84	74.3	50.8
ExtraTrees	0.68	4.42	74.8	51.5
Bagging	0.63	4.81	74.2	50.8

Table 5.A.30: Machine learning metrics using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PIP_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.61	3.45	83.1	56.9
GPR	0.63	3.38	84.9	64.1
ANN	0.51	4.15	79.4	55.0
SVM	0.60	3.53	87.1	67.8
PLS	0.56	3.68	83.4	57.6
RF	0.59	3.58	88.1	67.5
ExtraTrees	0.62	3.46	87.8	67.8
Bagging	0.60	3.57	88.1	66.1

Table 5.A.31: Machine learning metrics using descriptors calculated using TPSS/def2-TZVP for the ***ligands_CSD_PYR_set_TSSig_NoTS*** dataset.

Model	R^2	RMSE	% within 4.0	% within 2.0
MLR	0.58	4.83	72.6	41.9
GPR	0.63	4.53	77.2	50.3
ANN	0.54	5.28	66.7	41.2
SVM	0.69	4.19	80.8	56.7
PLS	0.58	4.83	73.0	45.4
RF	0.65	4.39	78.7	54.8
ExtraTrees	0.70	4.09	80.4	55.6
Bagging	0.65	4.39	78.8	54.4

Conclusions and Future Work

In conclusion, this work identified the need for a high-throughput computational method for the identification and prediction of ligand and catalyst activity for organometallic catalysis. Especially for the identification of novel ligands for base metal-catalysed reactions. The copper-catalysed Ullmann-Goldberg reaction was chosen as a test reaction to both develop and test the new methodology.

A new high-throughput computational methodology was developed to rapidly predict activation energies for a large number of ligands. Literature ligands for the Ullmann-Goldberg reaction were mined from the Reaxys database and used to test the viability of the workflow. Organometallic complexes of interest can be automatically generated and used to calculate the activation energy of the reaction of interest. The workflow predicts activation energies within 3.9 $\mathrm{kcal\,mol^{-1}}$ when compared to high-level coupled cluster methods, within the same time scale as a similar high-throughput experimental screen.

To enable the computational discovery of ligands for organometallic catalysis a new structural database, CatSD, was created to enable searching the Cambridge Structural Database via the CCDC's CrossMiner software. Ligands can be identified using a 3D structural query, called a catalophore, to define key structural features and search the CSD for similar molecules. The features included in CSD-CrossMiner were expanded to provide a feature set suitable for catalysis. A new set of coordinating atom features was developed to enable the definition of the coordinating environment of the metal. Currently, this method can only be applied to organic structures in the CSD. Further development could include expanding the applicability to organometallic structures in the CSD as well as generating databases of commercially available organometallic ligands to allow for computational screening in process chemistry and ligand selection.

CatSD was used to identify >10,000 ligands from the CSD for the Ullmann-Goldberg reaction. The activation energies for each ligand were tested using the developed workflow and several were chosen and tested experimentally. No conclusive trend was observed between activation energy and experimental yield. The deactivation pathways of these ligands were also explored again showing no conclusive trend between deactivation energy and experimental yield. We conclude that the reaction yield is not solely determined by ligand properties and is a combination of both ligand properties, deactivation pathways and more complex mechanistic aspects. This is supported by the literature which is still undecided on the full reaction mechanism for the Ullmann-Goldberg reaction as well as the lack of understanding of ligand properties and their effect on reactivity.

Machine learning was used to identify key ligand properties affecting the activation energy and predict activation energies without the need to calculate the transition state structure. Eight machine-learning models were tested on the four datasets generated from the high-throughput screening. A set of descriptors was developed to describe both the steric and electronic properties of the complexes using only the output of the workflow. ExtraTrees-based models were found to be the best at predicting activation energies with RMSEs approaching or surpassing the error of the calculated activation energies (3.9 $\text{kcal}\,\text{mol}^{-1}$). The machine learning models are an effective secondary tool for the identification of incorrect structures via the identification of outliers.

It was also demonstrated that descriptor sets excluding the transition states are able to predict the activation energies with similar accuracy. While transition states still need to be calculated to generate the models, predictions for new ligands are possible without the need to calculate transition state structures, reducing computational costs significantly.

Follow-up work to this project may include expanding CatSD to apply to organometallic structures within the CSD, creation of new databases containing commercially available ligands for process screening, expansion to open-shell complexes or converting the workflow presented in this workflow into a suite of software to make it readily accessible to the wider chemical community. We hope the methodologies presented in this work will provide a basis for the uptake and development of computational screening for organometallic catalysis. The tools presented are applicable to a large range of chemical applications from ligand discovery, process development and exploration of mechanistic pathways both industrially and in research.

References

(1) Atkins, P.; Overton, T., *Shriver and Atkins' Inorganic Chemistry*; OUP Oxford: 2010.

(2) De Vries, J. G.; Jackson, S. D. *Catal. Sci. Technol.* **2012**, *2*, 2009–2009.

(3) Neese, F. *WIREs Computational Molecular Science* **2018**, *8*, e1327.

(4) Sperger, T.; Sanhueza, I. A.; Schoenebeck, F. *Accounts of Chemical Research* **2016**, *49*, PMID: 27171796, 1311–1319.

(5) Poree, C.; Schoenebeck, F. *Accounts of Chemical Research* **2017**, *50*, PMID: 28945392, 605–608.

(6) JÓNSSON, H.; MILLS, G.; JACOBSEN, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, 1998, pp 385–404.

(7) Zimmerman, P. M. *The Journal of Chemical Physics* **2013**, *138*, 184102.

(8) Nieves-Quinones, Y.; Singleton, D. A. *Journal of the American Chemical Society* **2016**, *138*, PMID: 27794598, 15167–15176.

(9) Riplinger, C.; Neese, F. *The Journal of Chemical Physics* **2013**, *138*, 034106.

(10) Houk, K. N.; Cheong, P. H.-Y. *Nature* **2008**, *455*, nature07368[PII], 309–313.

(11) Ianni, J. C. et al. *Angewandte Chemie International Edition* **2006**, *45*, 5502–5505.

(12) Tantillo, D. J. *Angewandte Chemie International Edition* **2009**, *48*, 31–32.

(13) Occhipinti, G.; Koudriavtsev, V.; Törnroos, K. W.; Jensen, V. R. *Dalton Trans.* **2014**, *43*, 11106–11117.

(14) Holland, P. L. *Accounts of Chemical Research* **2015**, *48*, PMID: 25989357, 1696–1702.

(15) Freyschlag, C. G.; Madix, R. J. *Materials Today* **2011**, *14*, 134–142.

(16) Umile, T. P., *Catalysis for Sustainability: Goals, Challenges, and Impacts*; CRC Press: 2015.

(17) Bauer, E. B. In *Iron Catalysis II*, Bauer, E., Ed.; Springer International Publishing: Cham, 2015, pp 1–18.

(18) Kochi, J. K. *Accounts of Chemical Research* **1974**, *7*, 351–360.

(19) Li, P.; Merz, K. M. *Chemical Reviews* **2017**, *117*, PMID: 28045509, 1564–1686.

(20) Vogiatzis, K. D. et al. *Chemical Reviews* **2019**, *119*, PMID: 30376310, 2453–2523.

(21) Durand, D. J.; Fey, N. *Chemical Reviews* **2019**, *119*, PMID: 30802036, 6561–6594.

(22) Foscato, M.; Jensen, V. R. *ACS Catalysis* **2020**, *10*, 2354–2377.

(23) Nandy, A. et al. *Chemical Reviews* **2021**, *121*, PMID: 34260198, 9927–10000.

(24) Falivene, L. et al. *Nature Chemistry* **2019**, *11*, 872–879.

(25) Hopmann, K. H. *International Journal of Quantum Chemistry* **2015**, *115*, 1232–1249.

(26) Fey, N. et al. *Chemistry A European Journal* **2006**, *12*, 291–302.

(27) Fey, N. et al. *Organometallics* **2008**, *27*, 1372–1383.

(28) Ianni, J. C. et al. *Angewandte Chemie International Edition* **2006**, *45*, 5502–5505.

(29) Lipkowitz, K. B.; D'Hue, C. A.; Sakamoto, T.; Stack, J. N. *Journal of the American Chemical Society* **2002**, *124*, PMID: 12440925, 14255–14267.

(30) Knowles, R. R.; Jacobsen, E. N. *Proceedings of the National Academy of Sciences* **2010**, *107*, 20678–20685.

(31) Orlandi, M. et al. *Journal of the American Chemical Society* **2017**, *139*, PMID: 28475315, 6803–6806.

(32) Deeth, R. J.; Foulis, D. L.; Williams-Hubbard, B. J. *Dalton Trans.* **2003**, 3949–3955.

(33) Gensch, T. et al. *Journal of the American Chemical Society* **2022**, *144*, PMID: 35020383, 1205–1217.

(34) Tolman, C. A. *Chemical Reviews* **1977**, *77*, 313–348.

(35) Jover, J. et al. *Organometallics* **2010**, *29*, 6245–6258.

(36) Jover, J. et al. *Organometallics* **2012**, *31*, PMID: 24882917, 5302–5306.

(37) Fey, N. et al. *Dalton Trans.* **2020**, DOI: `10.1039/D0DT01694B`.

(38) Poree, C.; Schoenebeck, F. *Accounts of Chemical Research* **2017**, *50*, PMID: 28945392, 605–608.

(39) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. *Journal of Chemical Theory and Computation* **2018**, *14*, PMID: 30095903, 5249–5261.

(40) Harper, K. C.; Sigman, M. S. *Science* **2011**, *333*, 1875–1878.

(41) Fristrup, P.; Tanner, D.; Norrby, P.-O. *Chirality* **2003**, *15*, 360–368.

(42) Limé, E. et al. *Journal of Chemical Theory and Computation* **2014**, *10*, PMID: 26580763, 2427–2435.

(43) Rasmussen, T.; Norrby, P.-O. *Journal of the American Chemical Society* **2003**, *125*, PMID: 12708865, 5130–5138.

(44) Rosales, A. R. et al. *Nature Catalysis* **2019**, *2*, 41–45.

(45) Patrascu, M. B. et al. From Desktop to Benchtop A Paradigm Shift in Asymmetric Synthesis, 2019.

(46) Amar, Y. et al. *Chem. Sci.* **2019**, *10*, 6697–6706.

(47) Janet, J. P. et al. *Accounts of Chemical Research* **2021**, *54*, PMID: 33480674, 532–545.

(48) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. *Science* **2015**, *347*, 737–743.

(49) Wang, Y. et al. *Nucleic Acids Research* **2011**, *40*, D400–D412.

(50) Zahrt, A. F. et al. *Science* **2019**, *363*, DOI: 10.1126/science.aau5631.

(51) Janet, J. P.; Kulik, H. J. *Chem. Sci.* **2017**, *8*, 5137–5152.

(52) Janet, J. P.; Kulik, H. J. *The Journal of Physical Chemistry A* **2017**, *121*, PMID: 29095620, 8939–8954.

(53) Ioannidis, E. I.; Kulik, H. J. *The Journal of Physical Chemistry A* **2017**, *121*, PMID: 28059518, 874–884.

(54) Arunachalam, N. et al. Ligand additivity relationships enable efficient exploration of transition metal chemical space, 2022.

(55) Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. *JACS Au* **2022**, *2*, 1200–1213.

(56) Duan, C. et al. Exploiting Ligand Additivity for Transferable Machine Learning of Multireference Character Across Known Transition Metal Complex Ligands, 2022.

(57) Ahneman, D. T. et al. *Science* **2018**, *360*, 186–190.

(58) Das, M.; Sharma, P.; Sunoj, R. B. *The Journal of Chemical Physics* **2022**, *156*, 114303.

(59) Tian, H.; Rangarajan, S. *Journal of Chemical Theory and Computation* **2019**, *15*, PMID: 31419114, 5588–5600.

(60) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions, 2014.

(61) Gómez-Bombarelli, R. et al. *ACS Central Science* **2018**, *4*, PMID: 29532027, 268–276.

(62) Mennen, S. M. et al. *Organic Process Research & Development* **2019**, *23*, 1213–1242.

(63) Tian, J.; Wang, G.; Qi, Z.-H.; Ma, J. *ACS Omega* **2020**, *5*, PMID: 32905323, 21385–21391.

(64) Ullmann, F.; Bielecki, J. *Berichte der deutschen chemischen Gesellschaft* **1901**, *34*, 2174–2185.

(65) Ullmann, F. *Berichte der deutschen chemischen Gesellschaft* **1903**, *36*, 2382–2384.

(66) Ullmann, F.; Sponagel, P. *Berichte der deutschen chemischen Gesellschaft* **1905**, *38*, 2211–2212.

(67) Goldberg, I. *Berichte der deutschen chemischen Gesellschaft* **1906**, *39*, 1691–1692.

(68) Hurtley, W. R. H. *J. Chem. Soc.* **1929**, 1870–1873.

(69) Lindley, J. *Tetrahedron* **1984**, *40*, 1433–1456.

(70) Sugar, J.; Musgrove, A. *Journal of Physical and Chemical Reference Data* **1990**, *19*, 527–616.

(71) Conry, R. R. In *Encyclopedia of Inorganic Chemistry*; John Wiley & Sons, Ltd: 2006.

(72) Weingarten, H. *The Journal of Organic Chemistry* **1964**, *29*, 3624–3626.

(73) Zhang, S.; Zhang, D.; Liebeskind, L. S. *The Journal of Organic Chemistry* **1997**, *62*, PMID: 11671553, 2312–2313.

(74) Marcoux, J.-F.; Doye, S.; Buchwald, S. L. *Journal of the American Chemical Society* **1997**, *119*, 10539–10540.

(75) Aalten, H. et al. *Tetrahedron* **1989**, *45*, 5565–5578.

(76) Kiyomori, A.; Marcoux, J.-F.; Buchwald, S. L. *Tetrahedron Letters* **1999**, *40*, 2657–2660.

(77) Kelkar, A. A.; Patil, N. M.; Chaudhari, R. V. *Tetrahedron Letters* **2002**, *43*, 7143–7146.

(78) Daly, S. et al. *Organometallics* **2008**, *27*, 3196–3202.

(79) Cristau, H.-J. et al. *Organic Letters* **2004**, *6*, PMID: 15012063, 913–916.

(80) Ouali, A.; Spindler, J.-F.; Jutand, A.; Taillefer, M. *Advanced Synthesis & Catalysis* **2007**, *349*, 1906–1916.

(81) Strieter, E. R.; Blackmond, D. G.; Buchwald, S. L. *Journal of the American Chemical Society* **2005**, *127*, PMID: 15783164, 4120–4121.

(82) Strieter, E. R.; Bhayana, B.; Buchwald, S. L. *Journal of the American Chemical Society* **2009**, *131*, PMID: 19072233, 78–88.

(83) Job, G. E.; Buchwald, S. L. *Organic Letters* **2002**, *4*, PMID: 12375923, 3703–3706.

(84) Chang, J. W. W. et al. *Tetrahedron Letters* **2008**, *49*, 2018–2022.

(85) Jones, G. O.; Liu, P.; Houk, K. N.; Buchwald, S. L. *Journal of the American Chemical Society* **2010**, *132*, 6205–6213.

(86) Yu, H.-Z.; Jiang, Y.-Y.; Fu, Y.; Liu, L. *Journal of the American Chemical Society* **2010**, *132*, PMID: 21133430, 18078–18091.

(87) Maiti, D.; Buchwald, S. L. *Journal of the American Chemical Society* **2009**, *131*, PMID: 19899753, 17423–17429.

(88) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

(89) Cole, J. C. et al. *Journal of Chemical Information and Modeling* **2018**, *58*, PMID: 29425456, 615–629.

(90) Bissantz, C.; Kuhn, B.; Stahl, M. *Journal of Medicinal Chemistry* **2010**, *53*, PMID: 20345171, 5061–5084.

(91) Korb, O. et al. *Journal of Medicinal Chemistry* **2016**, *59*, PMID: 26745458, 4257–4266.

(92) The Cambridge Structural Database (CSD).

(93) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

(94) Louie, J.; Hartwig, J. F. *Tetrahedron Letters* **1995**, *36*, 3609–3612.

(95) Guram, A. S.; Rennels, R. A.; Buchwald, S. L. *Angewandte Chemie International Edition in English* **1995**, *34*, 1348–1350.

(96) Ioannidis, E.; Gani, T.; Kulik, H. *Journal of computational chemistry* **2016**, *37*, DOI: 10.1002/jcc.24437.

(97) Burton, V. J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J. *Journal of the American Chemical Society* **1995**, *117*, 8407–8415.

(98) Rappe, A. K. et al. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(99) Tubert-Brohman, I.; Schmid, M.; Meuwly, M. *Journal of Chemical Theory and Computation* **2009**, *5*, PMID: 26610220, 530–539.

(100) Rappe, A. K.; Colwell, K. S.; Casewit, C. J. *Inorganic Chemistry* **1993**, *32*, 3438–3450.

(101) Janet, J. P. et al. *Inorganic Chemistry* **2019**, *58*, PMID: 30834738, 10592–10606.

(102) Eksterowicz, J. E.; Houk, K. N. *Chemical Reviews* **1993**, *93*, 2439–2461.

(103) Norrby, P.-O. *Journal of Molecular Structure: THEOCHEM* **2000**, *506*, 9–16.

(104) Hansen, E. et al. *Accounts of Chemical Research* **2016**, *49*, PMID: 27064579, 996–1005.

(105) Rosales, A. R. et al. *Journal of the American Chemical Society* **2020**, *142*, PMID: 32249569, 9700–9707.

(106) Christensen, A. S.; Kuba, T.; Cui, Q.; Elstner, M. *Chemical Reviews* **2016**, *116*, PMID: 27074247, 5301–5337.

(107) Weber, W.; Thiel, W. *Theoretical Chemistry Accounts* **2000**, *103*, 495–506.

(108) Kazaryan, A. et al. *Journal of chemical theory and computation* **2011**, *7*, 2189–2199.

(109) Stewart, J. J. *Journal of Molecular modeling* **2007**, *13*, 1173–1213.

(110) Kromann, J. C.; Welford, A.; Christensen, A. S.; Jensen, J. H. *ACS omega* **2018**, *3*, 4372–4377.

(111) Minenkov, Y.; Sharapa, D. I.; Cavallo, L. *Journal of Chemical Theory and Computation* **2018**, *14*, PMID: 29787256, 3428–3439.

(112) Grimme, S.; Bannwarth, C.; Shushkov, P. *Journal of Chemical Theory and Computation* **2017**, *13*, PMID: 28418654, 1989–2009.

(113) Bannwarth, C.; Ehlert, S.; Grimme, S. *Journal of Chemical Theory and Computation* **2019**, *15*, PMID: 30741547, 1652–1671.

(114) Bursch, M.; Neugebauer, H.; Grimme, S. *Angewandte Chemie International Edition* **2019**, *58*, 11078–11087.

(115) Rasmussen, M. H.; Jensen, J. H. *PeerJ Physical Chemistry* **2020**, *2*, e15.

(116) Sure, R.; Grimme, S. *Journal of Computational Chemistry* **2013**, *34*, 1672–1685.

(117) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. *The Journal of Chemical Physics* **2015**, *143*, 054107.

(118) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. *The Journal of Chemical Physics* **2018**, *148*, 064104.

(119) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(120) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *The Journal of Chemical Physics* **2021**, *154*, 064103.

(121) Parr, R.; Weitao, Y., *Density-Functional Theory of Atoms and Molecules*; International Series of Monographs on Chemistry; Oxford University Press: 1994.

(122) Kohn, W.; Sham, L. J. *Physical Review* **1965**, *140*, 1133–1138.

(123) Kohn, W. *International Journal of Quantum Chemistry* **1995**, *56*, 229–232.

(124) Bühl, M.; Kabrede, H. *Journal of Chemical Theory and Computation* **2006**, *2*, PMID: 26626836, 1282–1290.

(125) Waller, M. P.; Braun, H.; Hojdis, N.; Bühl, M. *Journal of Chemical Theory and Computation* **2007**, *3*, PMID: 26636215, 2234–2242.

(126) Bühl, M. et al. *Journal of Chemical Theory and Computation* **2008**, *4*, PMID: 26621431, 1449–1459.

(127) Husch, T.; Freitag, L.; Reiher, M. *Journal of Chemical Theory and Computation* **2018**, *14*, PMID: 29595973, 2456–2468.

(128) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *The Journal of Chemical Physics* **2008**, *129*, 121104.

(129) Grimme, S. *Journal of Computational Chemistry* **2004**, *25*, 1463–1473.

(130) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *The Journal of Chemical Physics* **2010**, *132*, 154104.

(131) Neese, F. *Coordination Chemistry Reviews* **2009**, *253*, Theory and Computing in Contemporary Coordination Chemistry, 526–563.

(132) Poli, R. *Chemical Reviews* **1996**, *96*, PMID: 11848825, 2135–2204.

(133) Poli, R.; Smith, K. M.; Harvey, J. N. *Coordination Chemistry Reviews* **2003**, *238-239*, Article (Academic Journal), 347–361.

(134) Deng, L.; Margl, P.; Ziegler, T. *Journal of the American Chemical Society* **1999**, *121*, 6479–6487.

(135) Glascoe, E. A.; Sawyer, K. R.; Shanoski, J. E.; Harris, C. B. *Journal of Physical Chemistry C* **2007**, *111*, 8789–8795.

(136) Casitas, A.; Krause, H.; Goddard, R.; Fürstner, A. *Angew Chem Int Ed Engl* **2014**, *54*, 1521–1526.

(137) Harvey, J. N. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2006**, *102*, 203–226.

(138) Balasubramanian, K. *The Journal of Physical Chemistry* **1989**, *93*, 6585–6596.

(139) Reimers, J. R.; Cai, Z.-L.; Bilic, A.; Hush, N. S. *Annals of the New York Academy of Sciences* **2003**, *1006*, 235–251.

(140) Schneider, N. et al. *Chemistry A European Journal* **2009**, *15*, 11515–11529.

(141) Tlenkopatchev, M.; Fomine, S. *Journal of Organometallic Chemistry* **2001**, *630*, 157–168.

(142) Machura, B.; Jaworska, M.; Kruszynski, R. *Polyhedron* **2004**, *23*, 1819–1827.

(143) Zhao, Y.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2008**, *120*, 215–241.

(144) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(145) Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. *Journal of Physical Organic Chemistry* **2009**, *22*, 1127–1135.

(146) Martín, A.; Orpen, A. G. *Journal of the American Chemical Society* **1996**, *118*, 1464–1470.

(147) Fey, N.; Harris, S. E.; Harvey, J. N.; Orpen, A. G. *Journal of Chemical Information and Modeling* **2006**, *46*, PMID: 16563023, 912–929.

(148) Yang, K.; Zheng, J.; Zhao, Y.; Truhlar, D. G. *The Journal of Chemical Physics* **2010**, *132*, 164117.

(149) Becke, A. D.; Johnson, E. R. *The Journal of Chemical Physics* **2006**, *124*, 221101.

(150) Wheeler, S. E.; Houk, K. N. *Journal of Chemical Theory and Computation* **2010**, *6*, PMID: 20305831, 395–404.

(151) Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

(152) Whitten, J. L. *The Journal of Chemical Physics* **1973**, *58*, 4496–4501.

(153) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theoretical Chemistry Accounts* **1997**, *97*, 119–124.

(154) Kendall, R. A.; Früchtl, H. A. *Theoretical Chemistry Accounts* **1997**, *97*, 158–163.

(155) For Biotechnology Information, N. C. PubChem `https://pubchem.ncbi.nlm.nih.gov/` (accessed 05/05/2020).

(156) RSC ChemSpider `https://www.chemspider.com/` (accessed 05/05/2020).

(157) Elsevier Reaxys `https://reaxys.com/` (accessed 05/05/2020).

(158) CAS SciFinder `https://scifinder.cas.org/` (accessed 05/05/2020).

(159) Of Labour, U. S. D. OSHA Occupational Chemical Database `https://www.osha.gov/chemicaldata/` (accessed 05/05/2020).

(160) Institute, N. C. Chemical Identifier Resolver `https://cactus.nci.nih.gov/chemical/structure` (accessed 05/05/2020).

(161) O'Boyle, N. M. et al. *Journal of Cheminformatics* **2011**, *3*, 33.

(162) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *The Journal of Chemical Physics* **2000**, *113*, 9901–9904.

(163) Ásgeirsson, V. et al. *Journal of Chemical Theory and Computation* **2021**, *17*, PMID: 34275279, 4929–4945.

(164) Jacobson, L. D. et al. *Journal of Chemical Theory and Computation* **2017**, *13*, PMID: 28957627, 5780–5797.

(165) Ryu, H. et al. *Organometallics* **2018**, *37*, 3228–3239.

(166) Dunning, T. H.; Peterson, K. A. *The Journal of Chemical Physics* **2000**, *113*, 7799–7808.

(167) Semiempirical Extended Tight-Binding Program Package, `https://github.com/grimme-lab/xtb`, Accessed: March 2020.

(168) Frisch, M. J. et al. Gaussian 09 Revision D.01, Gaussian Inc. Wallingford CT, 2009.

(169) University of Leeds ARC Supercomputer, `https://arc.leeds.ac.uk/`, Accessed: April 2020.

(170) Bunnett, J. F.; Kim, J. K. *Journal of the American Chemical Society* **1970**, *92*, 7463–7464.

(171) Jenkins, C. L.; Kochi, J. K. *Journal of the American Chemical Society* **1972**, *94*, 856–865.

(172) Litvak, V. V.; Shein, S. M. *Zh. Org. Khim* **1974**, 2360–2366.

(173) Bacon, R. G. R.; Hill, H. A. O. *J. Chem. Soc.* **1964**, 1097–1107.

(174) Hey, D. H.; Waters, W. A. *Chemical Reviews* **1937**, *21*, 169–208.

(175) Zhang, C.; Tang, C.; Jiao, N. *Chem. Soc. Rev.* **2012**, *41*, 3464–3484.

(176) Schröder, K.; Konkolewicz, D.; Poli, R.; Matyjaszewski, K. *Organometallics* **2012**, *31*, 7994–7999.

(177) Bowman, W.; Heaney, H.; Smith, P. *Tetrahedron Letters* **1982**, *23*, 5093–5096.

(178) Bowman, W.; Heaney, H.; Smith, P. H. *Tetrahedron Letters* **1984**, *25*, 5821–5824.

(179) Fier, P. S.; Hartwig, J. F. *Journal of the American Chemical Society* **2012**, *134*, PMID: 22709145, 10795–10798.

(180) Abeywickrema, A. N.; Beckwith, A. L. J. *J. Chem. Soc., Chem. Commun.* **1986**, 464–465.

(181) Creutz, S. E.; Lotito, K. J.; Fu, G. C.; Peters, J. C. *Science* **2012**, *338*, 647–651.

(182) Tye, J. W. et al. *Journal of the American Chemical Society* **2008**, *130*, PMID: 18597458, 9971–9983.

(183) Van Allen, D. Mechanism of copper-catalyzed cross-coupling reactions. Ph.D. Thesis, University of Massachusetts, 2004.

(184) Turner, R. W.; Amma, E. L. *Journal of the American Chemical Society* **1963**, *85*, 4046–4047.

(185) Nicholls, B.; Whiting, M. C. *J. Chem. Soc.* **1959**, 551–556.

(186) Dargel, T. K.; Hertwig, R. H.; Koch, W. *Molecular Physics* **1999**, *96*, 583–591.

(187) Stibrany, R. T. et al. *Inorganic Chemistry* **2006**, *45*, PMID: 17112267, 9713–9720.

(188) Paine, A. J. *Journal of the American Chemical Society* **1987**, *109*, 1496–1502.

(189) Alemanga, A. et al. *Chemischer Informationsdienst* **1983**, *14*, no–no.

(190) Cohen, T.; Wood, J.; Dietz, A. G. *Tetrahedron Letters* **1974**, *15*, 3555–3558.

(191) Bethell, D.; Jenkins, I. L.; Quan, P. M. *J. Chem. Soc., Perkin Trans. 2* **1985**, 1789–1795.

(192) Malinakova, H. C. *Chemistry A European Journal* **2004**, *10*, 2636–2646.

(193) Sambiagio, C.; Marsden, S. P.; Blacker, A. J.; McGowan, P. C. *Chem. Soc. Rev.* **2014**, *43*, 3525–3550.

(194) Mansour, M. et al. *Chem. Commun.* **2008**, 6051–6053.

(195) Jiao, J. et al. *The Journal of Organic Chemistry* **2011**, *76*, PMID: 21261263, 1180–1183.

(196) Bacon, R. G. R.; Karim, A. *J. Chem. Soc., Perkin Trans. 1* **1973**, 278–280.

(197) Zhang, S.-L.; Liu, L.; Fu, Y.; Guo, Q.-X. *Organometallics* **2007**, *26*, 4546–4554.

(198) Lefèvre, G. et al. *Organometallics* **2012**, *31*, 7694–7707.

(199) Lefèvre, G. et al. *Organometallics* **2012**, *31*, 914–920.

(200) Franc, G. et al. *ChemCatChem* **2011**, *3*, 305–309.

(201) Giri, R.; Hartwig, J. F. *Journal of the American Chemical Society* **2010**, *132*, PMID: 20977264, 15860–15863.

(202) Ouali, A.; Taillefer, M.; Spindler, J.-F.; Jutand, A. *Organometallics* **2007**, *26*, 65–74.

(203) Sherborne, G. J. et al. *Chem. Sci.* **2017**, *8*, 7203–7210.

(204) Alvarez, S. *Dalton Trans.* **2013**, *42*, 8617–8636.

(205) Kabsch, W. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(206) Korb, O. et al. *Journal of Medicinal Chemistry* **2016**, *59*, PMID: 26745458, 4257–4266.

(207) Reich, H. J. Bordwell pKa Table (in DMSO) `https://www.chem.wisc.edu/areas/reich/pkatable/` (accessed 05/18/2020).

(208) The Periodic Table of the Elements.

(209) Mehmood, T.; Liland, K. H.; Snipen, L. G.; Sæbø, S. *Chemometrics and Intelligent Laboratory Systems* **2012**, *118*, 62–69.

(210) Kokabi, A.; Nasiri Mahd, Z.; Naghibi, Z. *Journal of Nanoparticle Research* **2021**, *23*, 157.

(211) Rupp, M. *International Journal of Quantum Chemistry* **2015**, *115*, 1058–1073.

(212) Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H. *The Journal of Physical Chemistry Letters* **2015**, *6*, PMID: 26722718, 3528–3533.

(213) Gómez-Bombarelli, R. et al. *Nature Materials* **2016**, *15*, 1120–1127.

(214) Byggmästar, J.; Nordlund, K.; Djurabekova, F. *Phys. Rev. Materials* **2020**, *4*, 093802.

(215) Shriver, D.; Atkins, P.; Langford, C. H. Inorganic Chemistry, W. H, 1990.

(216) GitHub - kjelljorner/morfeus: A Python package for calculating molecular features — github.com, [Accessed Oct-2021].

(217) Freixa, Z.; van Leeuwen, P. W. N. M. *Dalton Trans.* **2003**, 1890–1901.

(218) Bilbrey, J. A.; Kazez, A. H.; Locklin, J.; Allen, W. D. *Journal of Computational Chemistry* **2013**, *34*, 1189–1197.

(219) Hillier, A. C. et al. *Organometallics* **2003**, *22*, 4322–4326.

(220) Falivene, L. et al. *Organometallics* **2016**, *35*, 2286–2293.

(221) Verloop, A.; Hoogenstraaten, W.; Tipker, J. In *Drug Design*, Ariëns, E., Ed.; Medicinal Chemistry: A Series of Monographs, Vol. 11; Academic Press: Amsterdam, 1976, pp 165–207.

(222) Mulliken, R. S. *The Journal of Chemical Physics* **1955**, *23*, 1833–1840.

(223) Szabo, A.; Ostlund, N., *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover Books on Chemistry; Dover Publications: 1996.

(224) Mayer, I. *Chemical Physics Letters* **1983**, *97*, 270–274.

(225) Berquist, E. et al. Release of cclib version 1.7, 2021.

(226) Vermeeren, P.; Sun, X.; Bickelhaupt, F. M. *Scientific Reports* **2018**, *8*, 10729.

(227) Van Zeist, W.-J.; Visser, R.; Bickelhaupt, F. *Chemistry A European Journal* **2009**, *15*, 6112–6115.

(228) Mansson, R. A.; Welsh, A. H.; Fey, N.; Orpen, A. G. *Journal of Chemical Information and Modeling* **2006**, *46*, PMID: 17125199, 2591–2600.

(229) Stacklies, W. et al. *Bioinformatics* **2007**, *23*, 1164–1167.

(230) Akiba, T. et al. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

(231) Probst, P.; Boulesteix, A.-L. **2017**, DOI: `10.48550/ARXIV.1705.05654`.