Aalto University
School of Chemical
Engineering

Predictive Modelling of *N*-Heteroarene Cross-Couplings

Riku Petteri Holopainen

Advisor(s): Mikko Passiniemi (Orion Corporation), Juri Timonen (University of Helsinki)

Supervisor: Juha Siitonen (Aalto University)

Master's Programme in Chemical, Biochemical and Materials Engineering

Submitted for inspection, Espoo, July 2024.

Aalto University
School of Chemical
Engineering

Author Riku Holopainen

Title of thesis Predictive Modelling of *N*-Heteroarene Cross-Couplings

Degree programme Master's Programme in Chemical, Biochemical and Materials

Major Chemistry

Thesis supervisor Juha Siitonen

Thesis advisor(s) Mikko Passiniemi, Juri Timonen

Date 21.07.2024 **Number of pages** 111 **Language** English

Abstract

Palladium catalysed cross-couplings of dihalogenated heteroarenes have a natural tendency of proceeding selectively, retaining one halide. In general, site-selectivity is biased to occur at a halide adjacent to a heteroatom. While the synthesis of a single regioisomer is a relatively straightforward process, accessing substitution patterns disfavoured by cross-coupling continues to present a challenge.

Unconventional site-selectivity can be achieved through substrate- or catalyst-controlled approaches. Despite catalyst-controlled approaches being at the forefront of research, relatively little is known about the relationship between site-selectivity and a catalyst's structure. The development of novel reactions for unconventional selectivity has traditionally relied on trial-and-error based experimental work, occasionally supplemented by more systematic methods. However, these are still time and resource intensive processes. From early linear regression models to modern machine learning approaches, chemists have long sought to streamline this process through various computational methods by relating chemical structure to function.

This work sought to improve the development of novel reactions involving dihalogenated heteroarenes with the development of predictive models, along with the accompanying tools and workflows. More precisely, a classification model for predicting the regiochemical outcomes of various ligands in palladium catalysed Suzuki–Miyaura cross-couplings was developed. The initial model proved to accurately predict the outcomes of four experimentally verified out-of-sample ligands. After the second round of experiments, it became evident that selecting the correct base-solvent combination is as pivotal as choosing the appropriate ligand. The developed models should not only accelerate the discovery process for heteroarenes, but also the development of models accommodating other substrate classes.

Keywords cross-coupling, Suzuki–Miyaura, site selectivity, heterocycles, machine learning

Aalto-yliopisto
Kemian tekniikan korkeakoulu

Tekijä Riku Holopainen

Työn nimi Predictive Modelling of *N*-Heteroarene Cross-Couplings

Koulutusohjelma Master's Programme in Chemical, Biochemical and Materials

Pääaine Chemistry

Työn valvoja Juha Siitonen

Työn ohjaaja(t) Mikko Passiniemi, Juri Timonen

Päivämäärä 21.07.2024 **Sivumäärä** 111 **Kieli** englanti

Tiivistelmä

Palladiumkatalysoitujen dihalogenoitujen heteroareenien ristikytkennöillä on luontainen taipumus edetä selektiivisesti, säilyttäen yhden halidin. Paikkaselektiivisyys suosii yleisesti ottaen heteroatomin viereistä halidia. Vaikka näin yksittäisen regioisomeerin synteesi onkin suhteellisen suoraviivainen prosessi, epätavanomaisen paikkaselektiivisyyden saavuttaminen tuottaa edelleen haasteita.

Epätavanomainen paikkaselektiivisyys voidaan saavuttaa muun muassa substraatti- tai katalyyttiohjatuilla menetelmillä. Vaikka katalyyttiohjatut menetelmät ovat tutkimusten eturintamalla, paikkaselektiivisyyden ja katalyytin rakenteen välisestä suhteesta tiedetään loppupeleissä suhteellisen vähän. Uusien reaktioiden kehittäminen epätavanomaisen selektiivisyyden saavuttamiseksi on perinteisesti perustunut yritys ja erehdys pohjaiseen kokeelliseen työhön, jota on ajoittain täydennetty systemaattisemmilla menetelmillä. Nämä ovat kuitenkin aikaa ja resursseja vievä prosesseja. Varhaisista lineaarisista regressiomalleista nykyaikaisiin koneoppimismenetelmiin kemistit ovat pitkään pyrkineet yksinkertaistamaan tätä prosessia erilaisten laskentamenetelmien avulla yhdistämällä kemiallisen rakenteen funktioon.

Tällä työllä pyrittiin parantamaan dihalogenoituja heteroareeneja sisältävien uudenlaisten reaktioiden kehitystyötä ennustavien mallien sekä niihin liittyvien työkalujen ja työnkulkujen kehityksellä. Työssä luotiin luokittelumalli eri ligandien paikkaselektiivisyyden ennustamiseksi palladiumkatalysoiduissa Suzuki–Miyaura-ristikytkennässä. Alustava malli osoittautui ennustavan tarkasti neljän kokeellisesti varmennetun sekä otoksen ulkopuolista ligandia. Toisen koekierroksen jälkeen oli selvää, että oikean emäs- liuotinyhdistelmän valinta on yhtä keskeistä kuin sopivan ligandin valinta. Työssä luotujen mallien pitäisi nopeuttaa niin heteroareenien kuin muidenkin substraattiluokkien mallien löytö- ja kehitystyötä.

Avainsanat ristikytkentä, Suzuki–Miyaura, paikkaselektiivisyys, heterosyklit, koneoppiminen

Preface

The work presented in this thesis was carried out at the Orion Corporation in the department of Medicinal Chemistry between January and June of 2023. The writing and revising process had numerous obstacles in its way, but finally reached its conclusion in July of 2024.

I was initially approached by Juha Siitonen back in late 2022, who inquired me if I was interested in a thesis topic involving organic chemistry and machine learning. I chose to take this topic as both subjects have been long-standing interests of mine. As such, I would like to extend my gratitude to Juha and the Orion Corporation for offering me this great learning opportunity.

I would like to especially thank my advisors Mikko Passiniemi and Juri Timonen for their invaluable support in both topics of catalytic chemistry and data science. Without their help I would not be here. I am also thankful for the support provided by all my coworkers at the department of Medicinal Chemistry. Lastly, special thanks go to my friends and family for their continued support throughout my thesis work.

Espoo, 2024

Riku Holopainen

Table of contents

1 Introduction 1
2 Literature review 4
2.1 The Suzuki–Miyaura cross-coupling reaction 4
2.1.1 Overview 4
2.1.2 The general catalytic cycle 5
2.1.3 On conventional site-selectivity 9
2.1.4 Ligand-controlled site-selectivity 12
2.2 Data-driven chemistry 23
2.2.1 History of data-driven chemistry 23
2.2.2 Modelling and methodologies 27
2.2.3 Machine learning in cross-coupling reactions 36
3 Experimental section 42
3.1 Computational modelling 44
3.1.1 Substrate selection 44
3.1.2 Data acquisition and pre-processing 46
3.1.3 Palladium complex modelling and ligand parametrization 50
3.1.4 Feature selection and intercorrelation analysis 57
3.1.5 Classification modelling 64
3.2 Laboratory testing and re-training 69
3.2.1 Synthesis procedures 69
3.2.2 Purification and characterization 71
3.2.3 Re-training 72
4 Discussion 74
5 Conclusion 82
Supporting material 84
References 95

Symbols and abbreviations

AdaBoost	Adaptive boosting
AI	Artificial intelligence
BDE	Bond dissociation energy
CAS	Chemical abstracts service
COSY	Correlation spectroscopy
CREST	Conformer-rotamer ensemble sampling tool
DFT	Density functional theory
DPEPhos	Bis[(2-diphenylphosphino)phenyl] ether
DPPF	1,1'-Bis(diphenylphosphino)ferrocene
DT	Decision tree
DTBPF	1,1′-Bis(di-tert-butylphosphino)ferrocene
EA	Electron affinity
ESP	Electrostatic potential
FMO	Frontier molecular orbital
GPC	Gaussian process classifier
HMBC	Heteronuclear multiple-bond correlation spectroscopy
HOMO	Highest occupied molecular orbital
HPLC	High-performance liquid chromatography
HSQC	Heteronuclear single quantum coherence spectroscopy
HTS	High-throughput screening
IP	Ionization potential
IPR	1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene
IR	Infrared
KRAKEN	Kolossal viRtual dAtabase for moleKular dEscriptors of orgaNophosphorus ligands
LCMS	Liquid chromatography–mass spectrometry
LC-MS	Liquid chromatography-mass spectrometry
LFER	Linear free energy relationships
LLM	Large language model
LUMO	Lowest occupied molecular orbital
MESP	Molecular electrostatic potential

MFF	Multiple fingerprint feature
ML	Machine learning
MLP	Multi-layer perceptron
MLR	Multivariate linear regression
MORFEUS	Molecular features for machine learning
NBO	Natural bond orbital
NHC	N-heterocyclic carbene
NMR	Nuclear magnetic resonance
NN	Neural network
NOESY	Nuclear Overhauser effect spectroscopy
PCA	Principle component analysis
PEPPSI	Pyridine-enhanced precatalyst preparation stabilization and initiation
PPh3	Triphenylphosphine
PTA	1,3,5-triaza-7-phosphaadamantane
PtBu3	Tri-tert-butylphosphine
QPhos	1,2,3,4,5-Pentaphenyl-1'-(di-tert-butylphosphino)ferrocene
QSAR	Quantitative structure-activity relationships
QSPR	Quantitative structure-property relationships
RBF	Radial basis function
RF	Random forest
SASA	Solvent-accessible surface area
SMCC	Suzuki—Miyaura cross-coupling
SMILES	Simplified molecular input line entry specification
SVC	Support vector classification
SVM	Support vector machine
TFP	Tri(2-furyl)phosphine
XGBoost	Extreme gradient boosting
xTB	Extended tight binding

1 Introduction

Palladium catalysed cross-coupling reactions of dihalogenated heteroarenes have a natural tendency of proceeding selectively. Heteroarenes containing multiple halides serve as synthetically useful structural templates for increasing molecular complexity, as cross-coupling commonly occurs with predictable selectivity to give a product that retains one halide.[1] Halides, especially chlorides, are valuable structural features of many drug-like molecules. Of all the small-molecule drugs approved by the United States' Food and Drug Administration from 1994 to 2012, approximately 15% of these were chlorinated.[2]

In general, site-selectivity is biased to occur at halogens adjacent to a heteroatom, enabled by the intrinsic relative electrophilicity of the halides, as is the case in strongly polarized systems of many dihalogenated heteroarenes.[3] This enables a straightforward synthesis of a single regioisomer, but trying to access target molecules with substitution patterns disfavoured by cross-coupling remains a challenge.[1]

Unconventional site-selectivity can be achieved through substrate-controlled or catalyst-controlled approaches. In a substrate-controlled approach, substrates that bear non-identical halides are used to force site-selectivity. This is accomplished by taking advantage of the disparities in reactivity between the two halides.[4, 5] Cross-coupling takes place at the heavier halide, even in the case where a lighter halide is at a more favourable position to a heteroatom. However, greater synthetic effort is required to prepare dihalogenated heteroarenes with mixed halides compared to identical ones, which translates to increased costs.[1] Hence, strategies enabling site-selective reactivity without modifying the substrate are highly desirable.[6]

In a catalyst-controlled approach unconventional site-selectivity is achieved by changing the catalyst's structure, *ergo*, the ligands.[1] In contrast to a substrate-controlled approach, substrates bearing identical halides are used instead. This in turn makes site-selectivity a great deal more laborious task to achieve. Electronic and steric effects as well as the presence of different directing groups in the reactants are all important factors that affect the final regiochemical outcome of the reaction.[4]

In the case of the Suzuki–Miyaura cross-coupling reaction, one of the main determining factors for site-selectivity appears to be the relative easiness of the oxidative addition. For instance, the C-Cl bond α to nitrogen in 3,5-dichloropyridazine requires less energy to break, suggesting that the C3 site is more reactive than its distal counterpart.[4]

Despite catalyst-controlled selectivity being at the forefront of other research areas, comparatively little is known about the relationship between cross-coupling site-selectivity and a catalyst's structure. In most cases, well established mechanistic rationales for conventional selectivity have been established, but in others the mechanism remains unexplained. Most literature reports agree on the regiochemical outcomes of palladium catalysed cross-coupling reactions for most classes of dihalogenated *N*-heteroarenes, but several examples reveal that changing the catalyst's structure can invert site-selectivity entirely.[1]

Developing catalyst system for unconventional selectivity is a challenge in itself, yet additional consideration must be taken to avoid diarylation, further complicating the development of highly selective catalyst systems.[4] Traditionally, the design and optimization of novel reactions has relied on trial-and-error based experimental work. These have been occasionally supplemented by more systematic methods, such as high-throughput screening.[7, 8] However, these are still relatively time and resource intensive processes. From early linear regression models to more modern machine learning approaches, chemists have sought to streamline this challenging process through various computational methods by relating chemical structure to function.[9]

A branch of chemistry that would absolutely benefit from the speed-up provided by these technologies is discovery-oriented chemistry. As such, medicinal chemistry falls into this category. While numerous tools have been developed in the past with the aim of accelerating pharmaceutical processes, many aspects of the drug discovery process could still benefit from further refinement.[10] This thesis sought to improve this process with the development of a predictive model and the requisite workflows for modelling dihalogenated *N*-heteroarenes. More precisely, a classification model for predicting the regiochemical outcomes of various ligands in palladium catalysed Suzuki–Miyaura cross-couplings was developed. The developed models should not only accelerate the discovery

process for heteroarenes, but also the development of models accommodating other substrate classes.

This thesis is divided into two main categories: a literature review, and an experimental section. The literature review examines the general theory and mechanistic rationales behind the palladium catalysed Suzuki–Miyaura cross-coupling reaction in addition to presenting the evolution, various methodologies, and applications of data-driven chemistry. A greater emphasis is placed on machine learning approaches in a later chapter. The experimental section discusses the developed transition metal complex modelling workflows and machine learning classification models. The regiochemical outcomes of various commercially available ligands are predicted and their results verified in laboratory. Lastly, the results and future outlook of the various aspects of this thesis are discussed in the discussion and conclusion.

2 Literature review

The literature review of this thesis is divided into two halves. The first half discusses the theory behind the Suzuki–Miyaura cross-coupling reaction, its mechanism, and how different ligands can affect regioselectivity. The second half goes over the history of data-driven chemistry, its various methodologies, and how machine learning has been utilized in predicting the results of different cross-coupling reactions.

2.1 The Suzuki–Miyaura cross-coupling reaction

2.1.1 Overview

Ever since its discovery in 1979 by Akira Suzuki and Norio Miyaura, the Suzuki–Miyaura cross-coupling (SMCC) reaction has witnessed the most significant growth among the cross-coupling reactions. It is indeed one of the most widely used reactions to obtain functionalized aromatics and biaryls.[11]

The SMCC reaction provides a general methodology for forming carbon-carbon bonds.[12] The reaction is performed in the presence of a base between an electrophilic aryl halide and a nucleophilic organoborane, catalysed by a transition metal complex.[13] Palladium is typically the metal of choice in SMCC reactions, although more novel catalysts based on iron, copper, and nickel have been developed due to their wider availability, abundance, and lower economic cost. In addition, nickel catalysts have shown some advantages in certain reactions with respect to palladium.[11] The general reaction scheme for a palladium catalysed SMCC reaction is presented below in **Figure 1**.

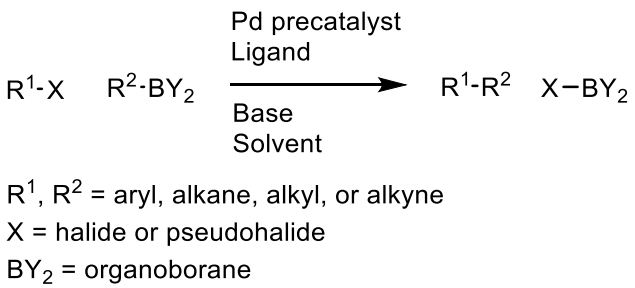


Figure 1. The general reaction scheme for a palladium catalysed Suzuki–Miyaura cross-coupling reaction.[11]

The main advantages of SMCC reactions compared to other cross-coupling reactions stem from the properties of the organoboron reagents. Organoboranes are readily available, air- and moisture-stable, more environmentally friendly, and less toxic compared to reactants such as organotin and organozinc. The inorganic boron-based byproduct can also be easily removed after the reaction.[12] Although the SMCC reaction always requires the presence of a base, the reaction can be performed under mild conditions as well as under an aqueous solvent and heterogenous conditions. Additionally, SMCC reactions are only slightly affected by steric hindrances and tend to proceed with high stereo- and regioselectivity. Although palladium as a catalyst might be expensive, only a small amount of it is required in the reaction, and it can be used in one-pot strategies.[11]

2.1.2 The general catalytic cycle

The mechanism of the SMCC reaction is best viewed from the perspective of the palladium catalyst. The reaction follows a well-defined catalytic cycle based on three basic steps: oxidative addition, transmetalation and reductive elimination.[11] Palladium catalysed SMCC's are generally based on the Pd(0)/Pd(II) cycle. The general catalytic cycle is presented below in **Figure 2**.

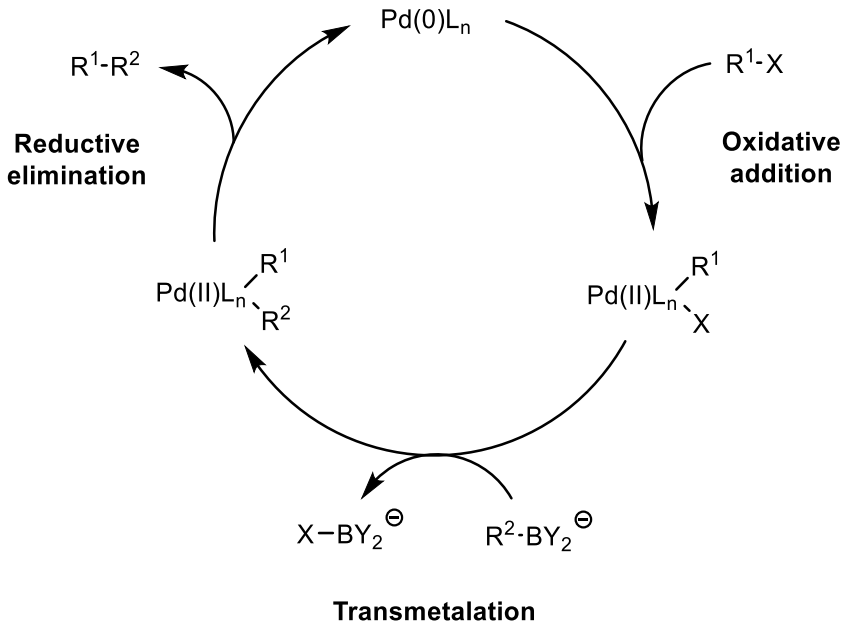


Figure 2. The general catalytic cycle for a palladium catalysed Suzuki–Miyaura cross-coupling reaction.[11]

The first step in the catalytic cycle is the oxidative addition of organic halides or other electrophiles to the Pd(0) complex.[11] This step can be reversible, but the equilibrium is expected to be displaced towards the product when using ligands with strong electron-donating abilities. Oxidative addition may follow two distinct mechanisms depending on various different factors like the solvent used and ligation state: an S_NAr-type mechanism or a three-center concerted mechanism.[1, 11, 14] The two proposed mechanisms for oxidative addition are presented below in **Figure 3**.

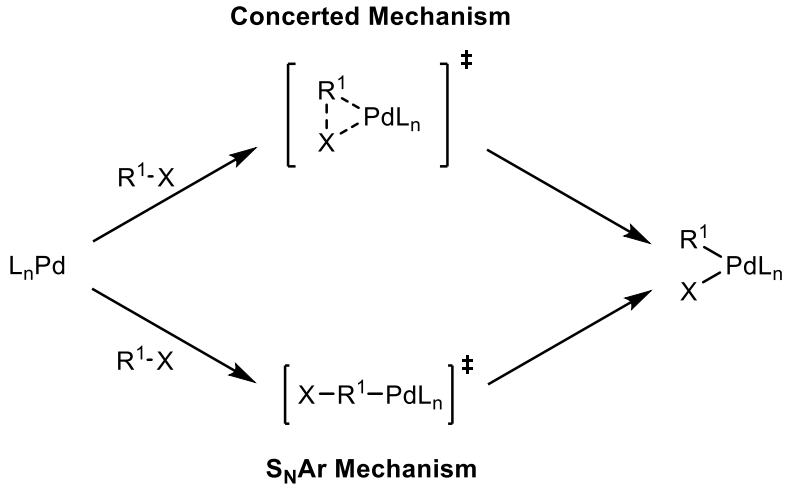


Figure 3. The two proposed reaction mechanisms for the oxidative addition step.[14]

The three-centered concerted mechanism entails the simultaneous forming of the Pd-R^1 and Pd-X bonds and the breaking of the R^1—X bond. This leads to the retention of configuration at a stereogenic carbon atom in the case of the chiral electrophile R^1—X and the relative cis disposition of the ligands R^1 and X after the oxidative addition.[11, 14] The second mechanism is an associative bimolecular process consisting of two steps. First, the carbon is attacked by the metal, expelling the anion X, and giving rise to a cationic species. Following this, both charged substances combine, and the palladium intermediate is formed. This second mechanism has been labelled as an S_NAr type reaction by analogy with the organic chemistry substitution reaction and results in the inversion of configuration if a stereogenic carbon center is involved.[14]

Regardless of the pathway, the bond between the heteroatom X and the organic R^1 group breaks and two new bonds are formed with the $PdL_n(0)$ metal complex. The oxidation state of the palladium is increased by two units and the Pd(II) intermediate is

formed.[11, 13] Initially a cis isomer, the palladium complex then undergoes isomerization to yield the more stable trans isomer via at least four concurrent pathways, two auto-catalytic and two solvent-assisted.[15] On the other hand, monodentate complexes may give trans isomers directly in lieu of *cis* isomers, avoiding the isomerization step entirely.[11] Isomerization from the cis isomer to the trans isomer is presented below in **Figure 4**.

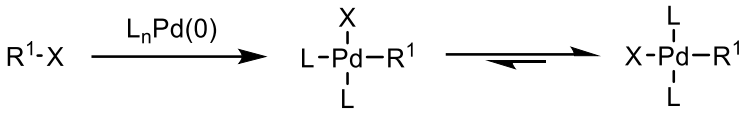


Figure 4. After the formation of the *cis*-palladium complex, the cis isomer isomerizes to the more stable trans isomer.[15, 16]

Oxidative addition is considered to be the rate-determining step in the catalytic cycle. Generally, it is the relative ease of the C—X bond breaking that determines the speed and/or the site-selectivity of the reaction.[11, 17] Relative reactivity of the halide or pseudohalide X decreases in the order of I > OTf > Br > Cl.[18]

The second step in the catalytic cycle is transmetalation. In this step the organic group R^2 is transferred from one metal to another, comparable to an exchange of ligands between two metals. This is the most characteristic step of the carbon-carbon bond forming cross-coupling reactions, as each cross-coupling reaction uses a different nucleophile.[11] For example, SMCC reactions use organoboranes whereas organozinc compounds are used in Negishi couplings and organotin compounds in Stille reactions.[13]

Contrary to the first and last step of the catalytic cycle, the mechanism behind transmetalation is not as well understood. The role of the external base was not initially clear, but it was common knowledge that it was necessary for this step, as without it the reaction would simply not occur.[11, 14] The requirement of the base leads to two hypothetical pathways, as is shown in **Figure 5**.

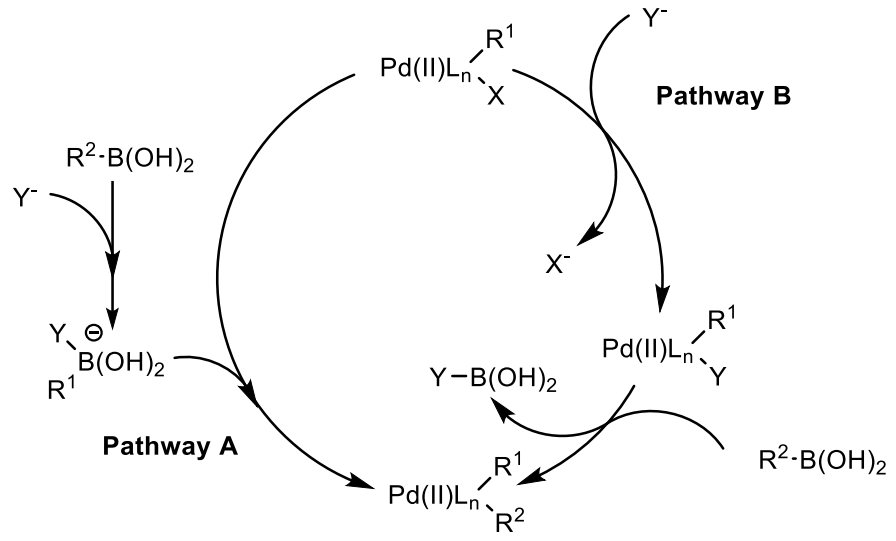


Figure 5. The two plausible pathways of the transmetalation step.[11]

In pathway A the base attacks the organoboron compound and forms an anionic ate complex, which then performs a nucleophilic attack on the palladium halide complex. The other mechanism, pathway B, takes place in two steps. In the first step a ligand substitution occurs between the halogen and the base. After that, the newly formed complex reacts with the neutral organoborane. Both pathways will lead to the same intermediate in the end, $Pd(II)L_n{-}R^1R^2$, which, however, has never been characterized.[11]

The last and the completing step in the catalytic cycle is reductive elimination, a reversal process of the oxidative addition step. This step comprises of the coupling of the organic groups R^1 and R^2 in cis position. Should the groups be in trans position to each other, a trans to cis isomerization reaction is required before this step. The generally accepted mechanism for this step is concerted and features a cyclic-three-coordinated transition state, similar to that of the oxidative addition, that results in the final C−C coupling. The oxidation state of the metal is concurrently reduced by two units and the Pd(0) active catalyst is regenerated. This step is usually irreversible.[11, 14] Additionally, like the oxidative addition step, this step is affected by different factors such as the characteristics of the ligands.[11]

Recent studies on the effects of ligands on this step show that bulky and electron-donating ligands favour this step. The steric hindrances of these types of ligands reduce the lifetime of the transitory $Pd(II)L_n{-}R^1R^2$ intermediate, increasing the rate of the

reaction. Consequently, this also prevents unwanted side reactions, such as β-elimination.[11, 19]

One last thing to note before going forward is that the previously described mechanisms only apply for mononuclear species of palladium catalysts. When using multinuclear catalysts, such as the trinuclear Pd_3X cluster, the cross-coupling reaction may instead catalyse through an entirely different mechanism, in which the transmetalation step occurs first.[20] The proposed new mechanism with Pd_3X is presented below in **Figure 6**.

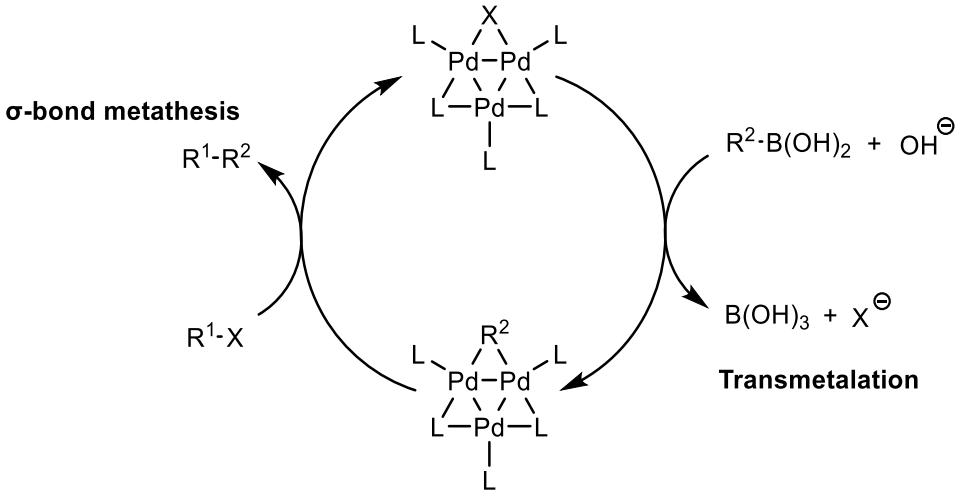


Figure 6. The proposed catalytic cycle for a Pd_3X-catalysed Suzuki–Miyaura cross-coupling reaction.[1, 20]

During the "new" catalytic cycle, Pd_3X will first react with phenylboronic acid to generate the intermediate Pd_3R^2.[20] The resulting intermediate will then react further with the aryl halide electrophile R^1—X via a σ-bond metathesis mechanism, giving the cross-coupled product and regenerating Pd_3X. In contrast to the "classic" SMCC catalytic cycle, the two coupling partners in this mechanism enter the catalytic cycle in reverse order, and there are no traditional redox steps involved.[1]

2.1.3 On conventional site-selectivity

For the purposes of this thesis, typical or conventional site-selectivity is defined as the regiochemical outcome currently reported in the vast majority of literature examples for a given substrate class, for example, 2,4-dichloropyridine and the substituted variations

thereof. Before delving deeper into how altering reaction conditions affect site-selectivity, the origins of conventional selectivity should be examined first.

For six-membered heteroarenes, cross-coupling typically takes place at a C—X bond adjacent to a heteroatom.[2, 3] In the case of dihalogenated pyridines, pyridazines and quinolines, this means the halide (or pseudohalide) α to a nitrogen atom. These sites are conventionally more reactive than their distal counterparts due to the polarity of the C—N causing the carbons to have a more positive charge. In turn, this makes them more reactive towards Pd(0). Additionally, C—X bonds α to a nitrogen are weaker because of the nitrogen's lone pair, which happens to be in the same plane as C—X.[1]

In 2006 Handy *et al.* demonstrated that cross-coupling site-selectivity could be predicted by analysing the ^{1}H nuclear magnetic resonance (NMR) chemical shifts δ_H of the parent heteroarene. More specifically, the proton that displays the largest chemical shift, or in other words, that is the most deshielded, is the typical site for cross-coupling as it is attached to the most electron-deficient carbon.[2, 21] While simple, the method is not foolproof, however. The method falls apart particularly in cases where $\Delta\delta_H$ was less than 0.3 ppm. Although the method proved accurate for several polysubstituted pyrroles, in the case of 3-arylpyrrole, where $\Delta\delta_H$ was just 0.02 ppm, site-selectivity could be switched from C2 to C4 by merely changing the solvent from DMF to an ethanol-toluene mixture.[22] Another example of the shortcomings of this method would be 2,4-dichloropyrimidine, where the most reactive site is C4 but the most deshielded proton is attached to C2.[1] Additionally, the method does not take into consideration steric factors, directing groups, or other interactions. It might also be unsuitable for heteroarene derivatives with more than two halides since the installation of the first substituent might influence the electronics of the heteroaromatic core.[21]

Later it would be Houk *et al.* that demonstrated through a series of computational studies that the site-selectivity in SMCC reactions would be mainly determined by the relative ease of oxidative addition.[1, 21] Site-selectivity in oxidative addition is controlled by two factors: the energy to distort a C—X bond to the transition-state geometry and the interaction between L_nPd and the aryl halide.[17] The easier the C—X bond is to distort into the transition-state geometry, the weaker the bond has to be. In other words, the smaller

the bonds bond dissociation energy (BDE) is, the more reactive the C—X bond usually is.[2] Although the oxidative addition step is generally considered to be the selectivity-determining step in the catalytic cycle, the importance of the transmetalation step for regioselectivity cannot be ignored.[23] It should be noted that the BDE is not the sole factor in determining the ease of oxidative addition. Many other structural features and especially the reaction conditions have a large influence in this step, which will be discussed in greater detail later.

This "distortion-interaction" or "activation-strain" model is able to rationalize many examples. However, when two C—X bonds have near identical BDEs, the interaction between L_nPd and the heteroaryl halide controls regioselectivity.[17, 22] The interaction energy ΔE_{int} between reactants is dominated by a favourable back-bonding secondary frontier molecular orbital (FMO) interaction as the bent L_nPd moiety approaches the C—X bond side-on. The distortion energy ΔE_{dist} contribution is relatively invariant when one type of halogen is involved although they note that in general BDE values are lower in six-membered rings compared to five-membered rings. BDE values were also lowered by the presence of a sulphur atom in the ring or when the halogen is an imonyl halide. The stabilising interaction energy term is dependent on the π^* LUMO coefficient, which is generally larger for positions adjacent to the ring's heteroatoms.[22] The distortion-interaction model for oxidative addition is shown below in **Figure 7**.

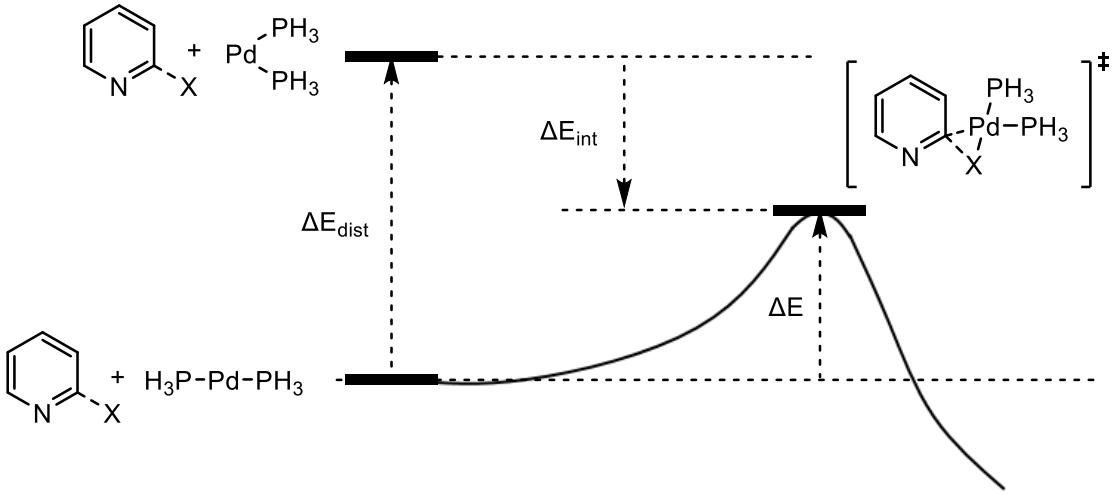


Figure 7. The distortion-interaction model for oxidative addition.[23, 24]

Further exceptions to the model were noted when the weaker C—X bond was on a carbocycle or when halides were adjacent to two different heteroatoms in a substituted arene. As stated previously, selectivity in these cases is better explained by a stronger attractive interaction between palladium and the more reactive site of the substrate.[1]

More recently Leitch *et al.* developed a model based on experimental results using $Pd(PCy_3)_2$ to predict substrate reactivity towards oxidative addition based on molecular descriptors. The model revealed that using a C—X bond's strength alone is a poor predictor of the rate of oxidative addition. Rather, the most important descriptor in this model was determined to be the electrostatic potential (ESP) at the carbon undergoing substitution, followed by the ESP at the atom adjacent to the reactive site.[1, 25] The model was able to predict the reactivity of diverse heteroaryl electrophiles by linking molecular structure to the predicted relative free energies of activation for any hypothetical substrate, with applicability well beyond the oxidative addition training set.[25]

Arguably, the biggest downfall of Leitch's model is that the parametrization was based only on results using $Pd(PCy_3)_2$ in a single solvent, tetrahydrofuran.[1] It is known that solvents play a role in activating precatalysts, stabilizing organometallic catalysts and reagents, and modulating the reactivity of acids and bases.[26] Thus, the calculated mechanism for oxidative addition can change with the use of solvation models, with palladium's coordination number, and with ligand identity. It is therefore reasonable to expect the transition states for C—X cleavage to look substantially different in cases where atypical site-selectivity is observed from that which has been hypothesized in Leitch's or Houk's models for conventional site-selectivity.[1] Regardless of the rationales behind these models, they do not, however, directly suggest explanations for the formation of the atypical product.

2.1.4 Ligand-controlled site-selectivity

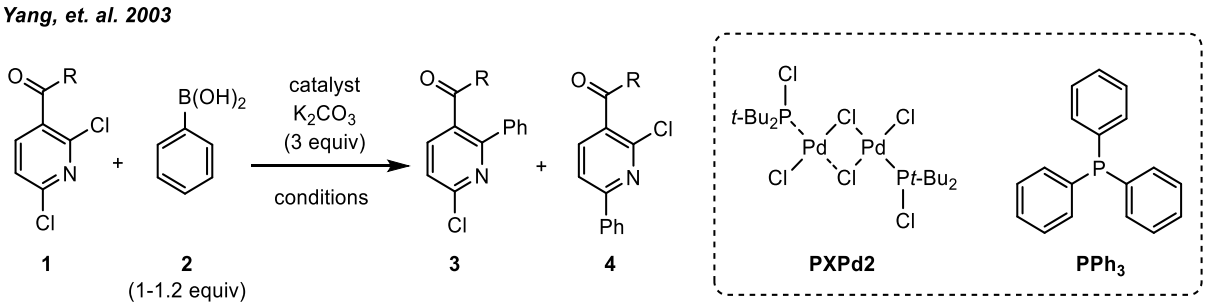
Until recently, unconventional site-selectivity had been for the most part achieved through various substrate-controlled strategies, which leverage the key differences in reactivity between various C—X bonds and directing groups.[6] However, multiple studies have shown that the combination of the solvent, the base, and the structure of the catalyst,

or more specifically, the ligands, can affect the yields and site-selectivity of SMCC reactions.[11] Ligand-controlled strategies, wherein ligands that bind exclusively to the transition metal can be used to alter the steric and electronic properties of the active catalyst, thus controlling the regiochemical outcomes regardless of the starting materials properties. In comparison to substrate-controlled strategies, ligand-controlled ones can increase the atom economy while simultaneously decreasing the step economy of the reactions by avoiding the later removal of undesired substituents. As a result, strategies that enable site-selective reactivity without modification of the substrate are highly desirable.[6] In addition, changes to the catalyst's speciation (from mono- to multinuclear) can have an equal or greater effect on site-selectivity compared to changes to the ligand environment at mononuclear palladium.[20, 27, 28] This thesis, however, will almost exclusively focus on the ligand's effects on site-selectivity of mononuclear palladium catalysed SMCC reactions.

Simple phosphine type ligands, such as triphenylphosphine (PPh_3), have been employed for a number of years in different palladium catalysed cross-coupling reactions. However, the need to overcome challenging substrate transformations under mild conditions and/or with higher selectivity has propagated the emergence of the field of ligand design. Deeper understanding of the steric and electronic properties of ligands as well as the binding properties of transition metals has led to the creation of well-defined organometallic complexes as highly effective precatalysts.[11] In recent years, unconventional site-selectivity has been achieved using new bulky. sterically hindered and electron-rich ligands. The above-mentioned phosphine ligands can dramatically improve the efficiency and selectivity of SMCC reactions as they enhance the rate of both the oxidative addition and reductive elimination processes.[29]

In 2003, Yang and co-workers reported on unconventional site-selectivity of different 2,6-dichloropyridine derivatives **1** with phenylboronic acid (**2**) (**Scheme 1**). Using PXPd2 as the catalyst, the team was able to selectively couple the aryl boron nucleophile at the atypical C2-site **3**, whereas a more traditional catalyst $Pd(PPh_3)_4$ afforded the more conventional C6 product **4**. The authors tested different base/solvent combinations and found K_2CO_3 and methanol to afford the highest selectivity.[30] This was one of the earliest examples demonstrating the possibility of ligand-controlled site-

selectivity of *N*-heteroarenes, although the reported selectivity's were not particularly high and were largely dependent on the directing group attached to the 2,6-dichloropyridine.[31]



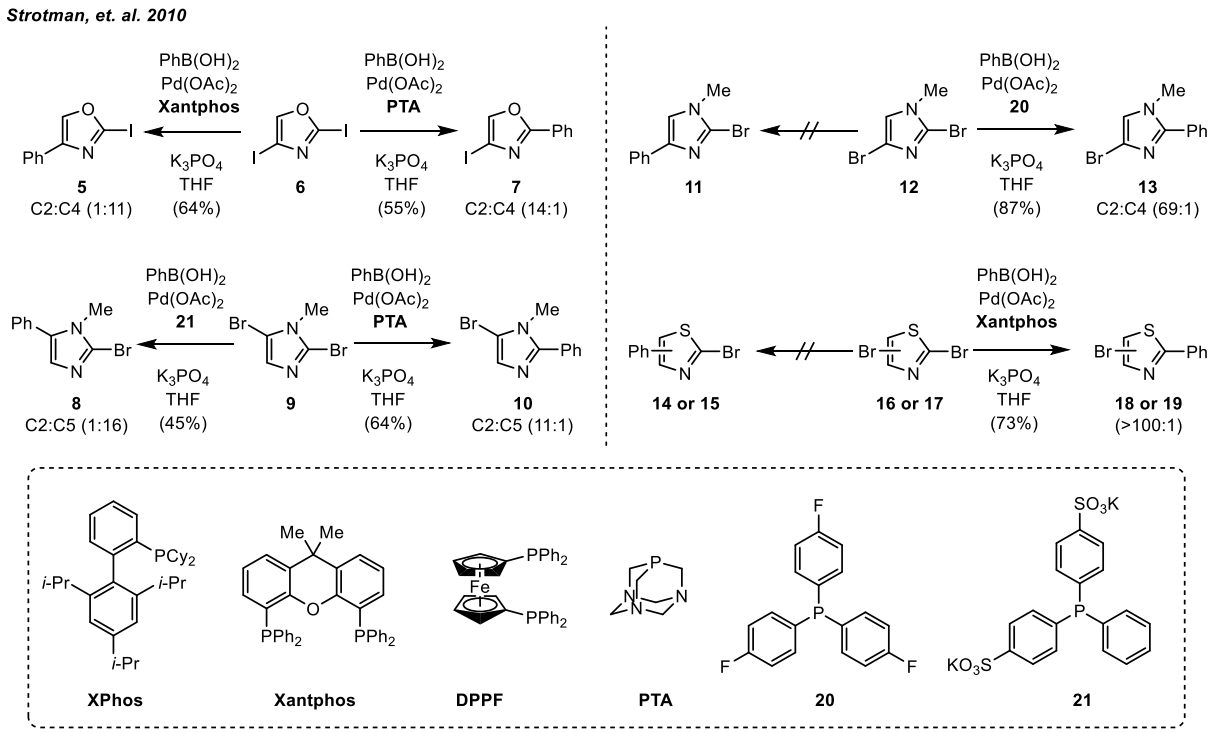
entry	R	catalyst (mol %)	conditions	3:4
1	OMe	$Pd(PPh_3)_4$ (5)	THF, reflux, 16h	1:5
2	OMe	PXPd2 (1)	MeOH, reflux, 30 min	2.5:1
3	$NHCH_2CH_2OPh$	PXPd2 (1)	MeOH, 55 °C, 1h	9:1

Scheme 1. Site-selective SMCC of 2,6-dichloropyridine derivatives. Yields were not indicated for entries 1 and 2. For entry 3 the yield of (**3**) is 61 %.[30, 31]

A few years later, Strotman *et al*. investigated the regioselective SMCC's of various dihaloimidazoles, dihalooxazoles, and dihalothiazoles, where several ligands were screened and identified to facilitate complementary selectivity outcomes (**Scheme 2**). They initially investigated the SMMC of 2,4-diiodooxazole (**6**) with phenylboronic acid and expected the more reactive site to be the C2-site based on Handy's method. Contrary to their prediction, most of the initially screened ligands favoured C4-selective coupling. Additionally, almost all the originally tested ligands offered poor site-selectivity and high levels of diarylation. However, Xantphos was found to be uniquely capable of mediating highly selective coupling at the C4 position **5**, as well as demonstrating high selectivity for monoarylation. To explore C2-selective coupling, the authors next surveyed around 200 achiral phosphine ligands. 1,3,5-triaza-7-phosphaadamantane (PTA) was identified as a ligand capable of high site-selectivity for the C2 position **7** and showed high selectivity towards monoarylation.[6, 30, 32]

With two sets of different conditions, the authors began their investigation of SMCC's with other commercially available dithiazoles, such as dibromo-1-methylimidazoles and

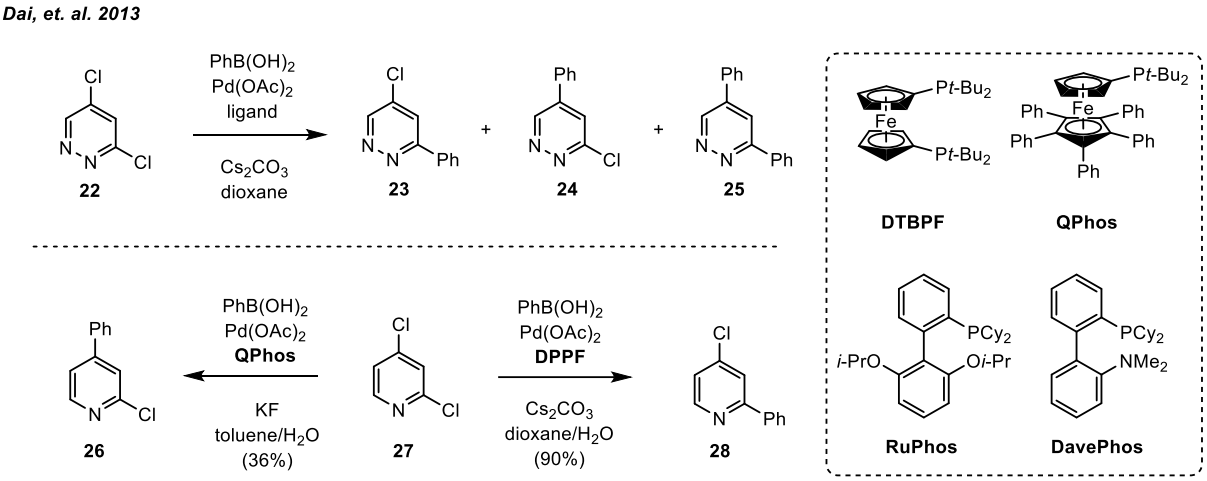
dibromothiazoles. Especially high C2 selectivity **8** was observed with the substrate 2,5-dibromo-1-methylimidazole (**9**) when PTA was employed as the ligand. However, when using a similar substrate, 2,4-dibromo-1-methylimidazole (**12**), none of the explored ligands afforded C4 position as the major product **11**. For dibromothiazoles **16** or **17**, Xantphos proved to be highly selective for C2 coupling **18** or **19**, but no ligand was identified that could selectively couple at either C5 or C4 positions **14** or **15**. The scope of the SMCC was extended to include a wide variety of electron-poor, electron-rich, and sterically hindered arylboronic acids.[32] Selectivity's with the acids followed similar trends to that observed with phenylboronic acid. The authors concluded that the regiochemical preference imparted by different ligands to be a cumulative effect of both steric and electronic effects.[6]



Scheme 2. Ligand-controlled site-selective SMCC's of various dihaloazoles and the ligands used to facilitate said reactions.[6, 32]

The first clear example of ligand-controlled inversion of site-selectivity was provided by Dai and co-workers back in 2013 (**Scheme 3**). Employing a popular bidentate ligand DPPF as the ligand showed exclusive coupling at the more conventional C3-site of 3,5-dichloropyridazine (**22**). However, when the authors switched to the bulky monodentate phosphine ligand QPhos instead, C5-selective coupling was achieved. Under optimized

conditions, Dai and co-workers were able to selectively couple numerous aryl, heteroaryl and alkenyl boron nucleophiles at the C5-site of 3,5-dichloropyridazine. QPhos was also shown to promote unconventional arylation using a different substrate entirely, 2,4-dichloropyridine (**27**), albeit with modest yield and selectivity.[4] Despite the limited yield, this is the first example of 2,4-dichloropyridine undergoing a C4 coupling in the absence of a directing group.[6]



entry	ligand	23	24	25
1	DPFF	82	0	18
2	Xantphos	69	12	19
3	PPh_3	76	0	24
4	QPhos	4	80	16
5	DTBPF	16	54	30
6	RuPhos	17	4	79
7	DavePhos	14	16	70
8	XPhos	10	15	75

Scheme 3. Ligand-controlled site-selective SMCC's of 3,5-dichloropyridazine and 2,4-dichloropyridine.[4, 6]

Whilst the authors did not provide a rationale for the unconventional selectivity at the time, they did note that chelation and electron density played a significant role in switching the site-selectivity. More specifically, electron deficient bidentate ligands favoured SMCC reactions at C3 over C5 whereas electron rich monodentate ligands favoured C5 over C3. Although DTBPF, an electron rich bidentate ligand, promoted reactions at C5 over C3 in a moderate ratio between mono- and diarylated product, the

team interpreted it as indicating that steric effects in addition to electronic effects could affect the site-selectivity of a SMCC reaction.[22] Furthermore, Buchwald's biaryl-based monophosphine ligands XPhos, DavePhos and RuPhos predominantly favoured diarylation of the product, even when only one equivalent of phenylboronic acid was used.[6]

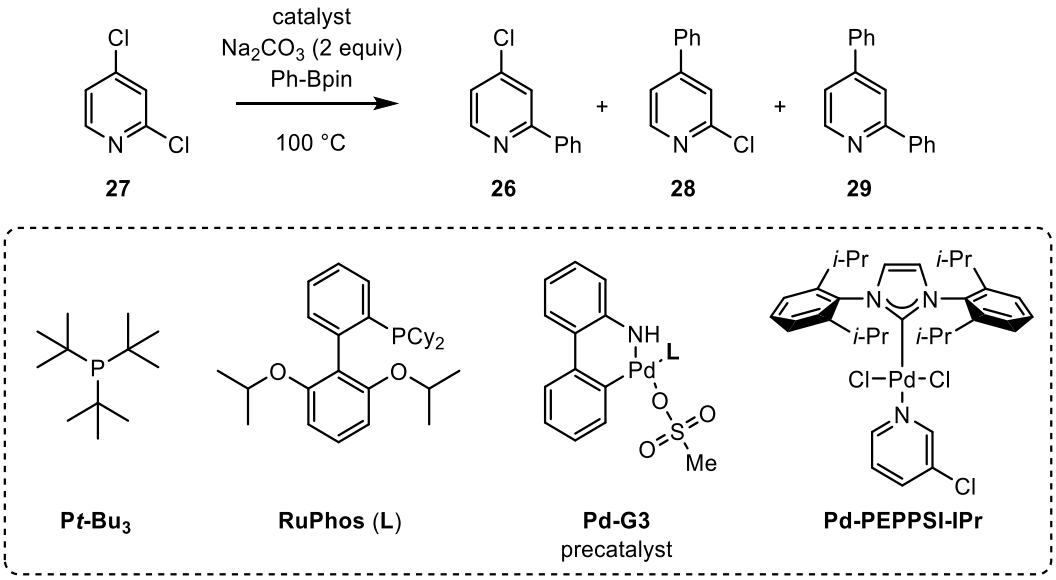
In addition to bulky phosphines, *N*-heterocyclic carbenes (NHCs) have emerged alongside as ligands that promote cross-coupling at a distal site of dihaloazines.[1] Sterically hindered NHCs have several desirable features when compared to tertiary and monodentate biaryl phosphines. For example, better oxidative, thermal, air and moisture stability when coordinated, distinct steric profiles, and better electronic properties as they can act as a strong σ-donor.[11]

Continuing the investigation of site-selective C4 couplings of 2,4-dichloropyridines (**27**), Yang and co-workers recently reported on a C4-selective SMCC using a sterically hindered NHC ligand 1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene (IPr) (**Scheme 4**). The authors began their work by extensively screening Pd-G3 precatalysts developed by the Buchwald group. In general, these palladium precatalysts from the corresponding family of biarylphosphine ligands are bulky, electron-rich ligands, which rapidly convert to the active PdL(0) species under mild reaction conditions. Compared to a more traditional precatalyst like $Pd(PPh_3)_4$, the Pd-G3 precatalysts gave better but moderate overall selectivity of C4:C2-coupled products.[33] A plausible mechanistic rationale is that the precatalysts containing bulky electron-rich ligands favour oxidative addition at the more sterically accessible C4 position. This logic would correspond with what Dai and coworkers observed during their studies.[6]

To further improve C4 coupling, Yang and co-workers began investigating "pyridine-enhanced precatalyst preparation stabilization and initiation" (PEPPSI) catalysts, which form highly stable organopalladium complexes containing bulky NHC ligands. Pd-PEPPSI-IPr was found to mediate the C4-arylated product, wherein site-selectivity was mainly associated with the steric bulk of the PEPPSI-IPr precatalyst, which inserted into the more sterically accessible C4 position of 2,4-dichloropyridine. However, a significant amount of diarylated product formed along with the desired C4-coupled product.[6] The

site-selectivity of these reactions seemed to largely depend on the solvent and base used. Notably, the hindered ligands PtBu3 and Pad2(n-Bu) effected modest C4 site-selectivity in a dioxane/water mixture. In this solvent system IPr and RuPhos gave the largest proportion of diarylated product, although switching the solvent to PEG400 decreased diarylation with IPr. Remarkably, a drastic improvement in C4-selectivity was observed by the addition of KI and NaOAc.[33]

Yang, et. al. 2020



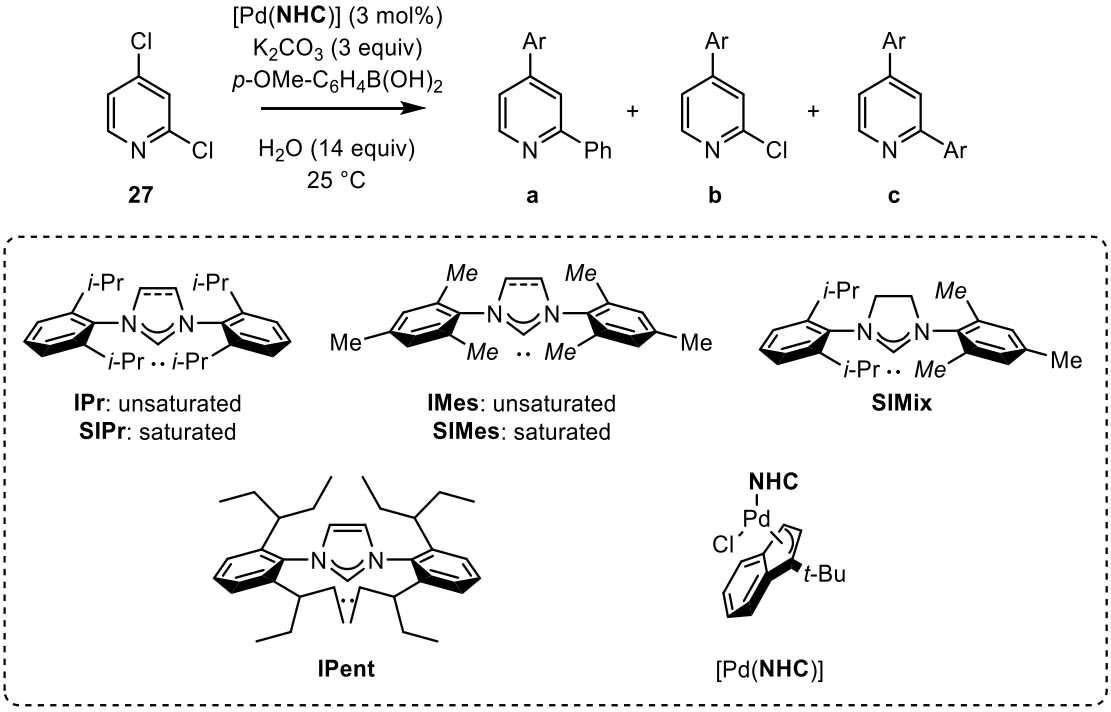
entry	catalyst	base	solvent	26:28:29
1	$Pd(PPh_3)_4$	Na_2CO_3	dioxane/H_2O (10:1)	2:80:18
2	RuPhos Pd-G3	Na_2CO_3	dioxane/H_2O (10:1)	15:37:48
3	$Pd(P\textit{t}\text{-}Bu_3)_2$	Na_2CO_3	dioxane/H_2O (10:1)	56:36:8
4	PEPPSI-IPr	Na_2CO_3	dioxane/H_2O (10:1)	37:8:55
5	PEPPSI-IPr	Na_2CO_3	toluene/H_2O (10:1)	61:24:15
6	PEPPSI-IPr	Na_2CO_3	PEG400	69:16:15
7	PEPPSI-IPr	NaOAc	PEG400	83:8:9
8	PEPPSI-IPr	NaOAc + Na_2CO_3	PEG400	99:1:0

Scheme 4. Ligand controlled SMCC's of 2,4-dichloropyridine.[1, 33]

Further exploration on the role of ligand sterics affecting the site-selectivity's of 2,4-dichloropyridine derivatives was conducted very recently by Norman *et al*. The authors screened various NHC ligands, both saturated and unsaturated, and in agreement with earlier reports, found that the steric bulk of the ligand appears to control site-selectivity (**Scheme 5**). In essence, the bulkier the NHC ligand, the greater the distal site-selectivity.

However, with increased ligand size comes a trade-off. The proportion of diarylated product increases concurrently with the C4:C2 ratio of monoarylated products. When using smaller ligands such as SIMes and IMes, much more of C2 site-selectivity was observed. Interestingly, the unsaturated ligand IMes was slightly more selective than its saturated counterpart, paralleling the trend seen with IPr versus SIPr. In a similar manner, ligands whose sterics were intermedial between SIMes and SIPr gave mixed selectivity. The use of the ligand IPent, which bears longer alkyl ortho-substituents, lead to notable diarylation. The authors concluded that IPr offered the best compromise between high site-selectivity and low diarylation, therefore maximizing the yield of the unconventional C4 product.[2]

Norman, et. al. 2022



entry	NHC	solvent	a	b	c
1	SIMes	benzene	24	35	3
2	IMes	benzene	23	40	3
3	SIMix	benzene	8	34	2
4	SIPr	benzene	9	64	8
5	IPr	benzene	8	70	6
6	IPr	THF	7	69	8
7	IPent	benzene	1	34	23

Scheme 5. Various NHC ligands and their reaction outcomes.[1]

Under optimized mild conditions, Norman and co-workers were successfully able to couple at the distal site of various 2,4-dichloropyridines and-quinolines as well as 3,5-dichloropyridazines. However, under these same room-temperature conditions, no reaction was observed in the absence of IPr. They further noted that under higher temperature conditions, the use of ligand is not always necessary to achieve high unconventional site-selectivity. In fact, notably high distal site-selectivity was observed in the cross-couplings of various dihaloheteroarenes. The authors suggest that the unprecedented selectivity observed in the aforementioned ligand-free "Jeffery-Heck" conditions may be a consequence of a multinuclear palladium species.[2]

Thus far multiple different studies have highlighted an unmistakeable relationship between ligand sterics and unconventional site-selectivity, but the mechanistic rationales behind these phenomena have not been discussed in detail. Analysis done on FMOs by Norman and co-workers suggests that this correlation may be explained by the differences in the orbital symmetries of 12-electron and 14-electron palladium species.[34]

For a traditional mononuclear species of palladium, oxidative addition is considered to be the site-selectivity determining step. As stated earlier, two competing mechanisms have been previously proposed: a three-centered concerted mechanism and a displacement type mechanism, which has also been referred to as an S_NAr-like mechanism, or a dissociative process. The preference for different mechanisms for different palladium species may arise from the differences in their orbital symmetries. The HOMO of a 14-electron PdL_2 has a π-type symmetry once it has been distorted into the required geometry for the forming of the pre-oxidative addition π-complex with the substrate. This π-symmetry enables the palladium to backbond with both the ipso and ortho carbons of the substrate via a nucleophilic displacement mechanism, but it does not directly interact with a halide. In comparison, the HOMO of a 12-electron PdL is a σ-type orbital, thus its symmetry is suited to the three-centered concerted mechanism. In this mechanism the palladium donates only to the ipso carbon, or in other words, to a single lobe of the substrate's LUMO. This allows the palladium to react at the site with a significantly larger LUMO coefficient, which can be, for example, a site distal to a nitrogen in a heteroarene.[34] The two different mechanisms of Norman and co-workers model system comprising of PhCl and $Pd(PMe_3)_n$ ($n = 1$ or 2) is presented in **Figure 8**.

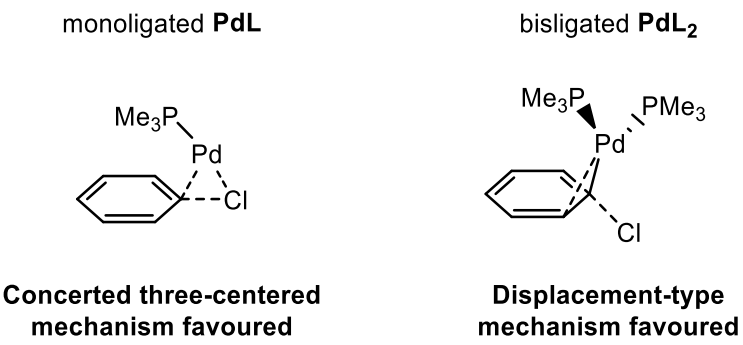


Figure 8. Monoligated Pd(PMe_3) and bisligated Pd(PMe_3)$_2$ displaying how PdL and PdL_2 may favour different mechanistic paths for oxidative addition.[1, 34]

The authors also remarked that the simple model ligands they used many not always be adequate for describing the behaviour of more complex ligands. In their case, however, their density functional theory (DFT) calculations with the simple model system proved to be consistent with those obtained using Pd(IPr) and 2,4-dichloropyridine.[34]

For 2,4-dichloropyridine, the percent contribution to the LUMO is substantially greater at the C4-site, accounting for 26% of the overall LUMO density whereas C2 only contributes around 8%. It is no wonder, then, why PdL and PdL_2 may exhibit different site-selectivity. As the σ-type HOMO of PdL can only donate into one ring atom of pyridine, the larger LUMO coefficient at C4 is a more important factor as it dictates the strength of the HOMO(Pd)/LUMO(substrate) overlap. In contrast, as the π-type HOMO of PdL_2 donates into two ring atoms of pyridine during a nucleophilic displacement mechanism, a strong orbital overlap can be achieved at the C2-site despite its smaller LUMO coefficient. For PdL_2, the selectivity for oxidative addition is mostly determined by other factors such as BDE and electrostatic potential.[34]

The authors calculated transition structures of monoligated Pd(IPr) suggests that the lowest energy pathway for reaction at either C2 or C4 involves the three-centered concerted transition state. For bisligated Pd(IPr), the reaction is predicted to proceed through the displacement-type mechanism at either site, with C2-site being favoured by this mechanism. Cross-couplings of other dihaloheteroarenes with Pd(IPr) also displayed a correlation between large differences in substrate LUMO coefficients and distal site-selectivity. Preferential reaction with Pd(IPr) at the C2-site was observed when two sites

had nearly identical LUMO coefficients. Additionally, Norman and co-workers calculated both the transition structures of mono- and bisligated Pd(P*t*-Bu$_3$) and Pd(IMes). Similar mechanistic observations to Pd(IPr) were made with mono- and bisligated Pd(P*t*-Bu$_3$), but the authors found P*t*-Bu$_3$ to be less C4-selective. This is likely due to phosphines having weaker σ-donicity in comparison to NHC ligands. Calculations with bisligated Pd(IMes) predicts that the displacement mechanism at the C2-site to be energetically competitive with the concerted mechanism for reaction at the C4-site using monoligated Pd(IMes). This prediction is consistent with experimental data as poor selectivity has been observed in the past with Pd/IMes-catalysed reactions. Past experiments also show that site-selectivity is skewed towards the C2-coupled product when higher concentrations of 2,4-dichloropyiridine are employed.[34]

Based on these highlighted works, it appears that accessing the 12-electron PdL, or the displacement-type mechanism, is critical to achieving unconventional site-selectivity. However, whereas traditionally the active catalyst has been obtained by mixing *in situ* an excess of the ligand with the Pd(0) source, PdL may not always be generated by such means as in solution more than one ligand would coordinate to the palladium.[11] Therefore, the use of a precatalyst or ligands that specifically promote oxidative addition at a low-coordinate species is necessary. Bulky and sterically hindered and strong σ-donor ligands appear to favour mono-coordination, while smaller ligands and better π-acceptors, such as PAr$_3$ and saturated NHCs, tend to favour the coordination of a second ligand. Computational studies have also demonstrated that solvent properties can affect the preferred mechanism, as polar solvents or an externally applied electric field seem to favour the displacement-type pathway.[1]

In addition to ligand and solvent properties, factors such as substrate concentration has been proven to affect the palladium's coordination number.[1] As such, the litany of factors that seem to affect the selectivity of SMCC's appears endless. Developing a deeper understanding of the conditions under which these mechanisms are relevant will undoubtedly bring ligand-controlled approaches to the forefront of the field of site-selective cross-couplings and open up the possibility for less-explored substrate classes.

2.2 Data-driven chemistry

2.2.1 History of data-driven chemistry

On the basis of the previous chapter, it should be clear by now that cross-coupling reactions, and for that matter, organic chemistry is filled with complex relationships and many unanswered questions. A longstanding approach that has sought to answer these questions is data-driven modelling.[35] To understand where we are today, we must first examine what different discoveries and developments in data-driven chemistry have led to this point in time. The concise timeline for the history of data-driven chemistry is presented below in **Figure 9**.

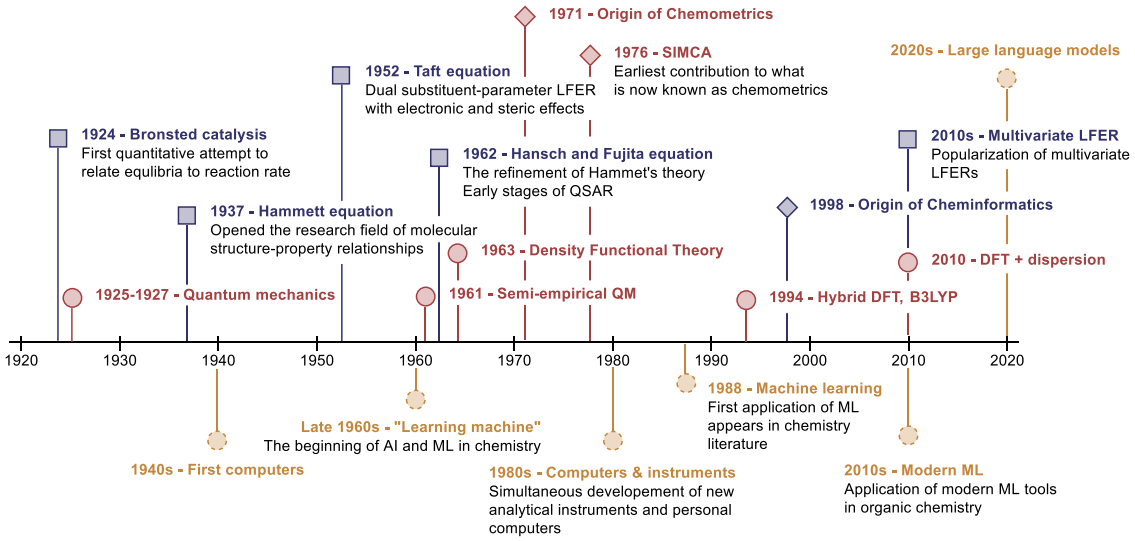


Figure 9. The timeline of the major developments in data-driven chemistry.[35, 36]

Many of the modern strategies are founded upon linear free energy relationships (LFERs), a method for relating reactivity with chemical structure. These are often represented by quantitative experimental parameters or descriptors. As the relationships are linear in energy, changes in a molecule's structure alters the Gibbs free energy in an additive manner. Several of the first parameters used in LFERs were derived from reaction equilibria, such as the Brønsted catalysis law. In 1924, Brønsted and co-workers derived the first quantitative relationship between equilibria and reaction rate by relating the ionization constant of acids to the rate of acid catalysed reactions via a sensitivity factor α. Hence, the dissociation of acids or bases could be utilized as a reference process linked to the outcomes of entirely different reactions.[35] Around the

same time frame, the mathematical foundational principles of quantum mechanics were established by Heisenberg, Schrödinger, and Dirac.[36]

In 1937, Hammett introduced an equation that can be considered to be the paradigmatic example of LFERs. In essence, the Hammett relationship correlates the substituents effects on resonance and induction by comparing the changes imparted by it relative to an unsubstituted benzoic acid. Numerous reactions involving substrates, reagents, or catalysts bearing substituted aromatics or other fragments has been found to correlate with the benzoic acid's acidity. Due to the versatility of the method, virtually every class of organic reaction has been analysed using the Hammett equation or its extended forms.[35]

Up until the 1950s, LFERs assumed a linear relationship between a single parameter and chemical reactivity. However, the simplicity of univariate LFERs limited the attainable information in more complex scenarios. The outcomes of reactions are more often than not dependant on a multitude of different factors that correlate nonlinearly.[35] In 1952, Taft reported a two-variable LFER that derived steric and electronic parameters from the rates of acid- and base-catalysed esterification/hydrolysis. Taft assumed acid-catalysed hydrolysis would be solely influenced by steric effects, whereas the base-catalysed variant would be influenced by both steric and electronic effects. While an oversimplification, essentially all chemical reactivity can be divided into steric and electronic effects.[35, 37] Building on Taft's multiparameter approaches, in 1962, Hansch, Fujita and co-workers would move the field of LFERs toward phenomena more relevant to pharmaceutical sciences and biochemistry by introducing correlations with partition coefficients. This advancement is considered to be the origin of quantitative structure-activity relationships (QSARs) and the foundation for the field of cheminformatics, a term that would be introduced several decades later.[35]

With the advent of computers and advancements in quantum chemical methods in the 1960s, namely the introduction of semi-empirical quantum mechanics in 1961 and DFT in 1963, LFERs were no longer limited to relying solely on experimentally derived parameters.[35, 36] Computational parameters offer numerous benefits over experimental ones, such as access to parameters with no experimental equivalent and the ability to

parametrize a structure before synthesis. Methods like DFT offer relatively good accuracy at low computational cost with single structures. However, when the computation of several structures is necessary, such as in the case of computing multiple conformers, computational costs may become a challenge. The sensitivity to the used model system used also proves to be a challenge with computational parameters as the functional set, solvent model and parametrization of a single conformer or several conformers all affect said parameters. The late 1960s also marks the emergence of artificial intelligence (AI) and machine learning (ML) in the form of "computerized learning machine" and other learning algorihms.[35]

The year 1971 would see the coining of the term “chemometrics”. In essence, chemometrics is the application of mathematical and statistical tools in chemistry. A more precise definition would be that it is a chemical discipline that applies mathematics, statistics, and formal logic to design optimal experimental procedures while providing maximum relevant information regarding chemical systems. Additionally, with the introduction of computers, LFERs would see a gradual progression from single variable methods to multivariate algorithms, ultimately leading to large-data chemometric methods.[35]

As computational power increased in the 1980s and 90s, *in silico* chemistry would significantly open up areas of possible chemical exploration, eventually leading to the field of cheminformatics. Cheminformatics can be described as a theoretical discipline of chemistry, complementary to quantum chemistry and force-field molecular modelling, with the focus on describing molecular structure in a format friendly for the use in statistical modelling, or ML. The field is primarily associated with QSARs or quantitative structure-property relationships (QSPRs) with a focus on drug-lead identification. Early QSAR models were predominantly based on multivariate regression with limited features. Limitations of these early approaches include sole suitability for linear modelling methods and being valid only to closely related compounds. In comparison, modern QSAR methods capable of handling large-scale data sets can make use of global models, which are trained on a wide range of compounds, even those lacking structural similarity. Generally, a cheminformatics model consists of a two-part process: converting molecules to feature vectors and then mapping of the said features to the property of

interest. Compared to other branches of computational chemistry, cheminformatics involves data processing that simply cannot be done without *in silico* mathematics. Additionally, cheminformatics is dependent on large scale data sets that cannot be compressed to standard mathematical models.[35]

Similar to chemometrics, cheminformatics is reliant on *in silico* mathematical, statistical, and ML methods in order to transform chemical data into processable information. While highly similar, the two fields should not be confused with each other. The primary distinction of chemometrics lies in its utilization of multivariate data obtained from instrumental analysis, where knowledge of the chemical structure is not always a necessity. Conversely, cheminformatics focuses on data generation directly from the description of the chemical structure. As such, cheminformatics can be considered to be a specific application of ML with an emphasis on modelling structure-property relationships for different molecules.[35]

In the 1990s, AI and ML based methods began gaining relevance in the field of chemometrics and cheminformatics, marking a pivotal point in the evolution of chemical analysis methodology. These ML tools found a wide range of different applications, ranging from molecular design to reaction prediction and optimization.[35, 38] Many of the early applications of ML relied on multivariate linear regression and other relatively simple statistical methods, often applied to comparatively small data sets.[10] Until very recently, the availability of the data has often been the largest limiting factor.[39] Despite the limitations, these early applications laid the groundwork for more sophisticated approaches that would emerge in the following decades.

Thus, we arrive at today where ML and AI are one of the most rapidly evolving topics in various frontiers of different industries, including medicinal chemistry and drug discovery.[10, 40] As computing power and the availability of data has increased significantly in the last couple of decades, it is no wonder that AI has seen a remarkable surge in popularity.[41, 42] ML has found a wide range of different applications in organic chemistry, ranging from molecular design to reaction prediction and optimization.[35, 38, 43] Furthermore, the utilization of ML has resulted in the identification of entirely new reactions and molecules that may not have been possible based on a chemist's reasoning

or intuition alone. Although these methods generate a wide array of different suggestions, most of which are, more often than not, still erroneous. It is still up to the chemist's judgement to choose the most reasonable outcomes.[35] More detailed examples of ML applications relevant to this thesis, cross-couplings in particular, will be reviewed in a later chapter.

2.2.2 Modelling and methodologies

The development of new reaction methodologies, particularly in organometallic catalysis, can prove to be a challenging task. Traditionally, reaction optimization has been achieved through exhaustive and trial-and-error-based reaction screenings, with the occasional application of systematic approaches, such as high-throughput screening (HTS).[7-9] Modern HTS is based on automated techniques with the goal of rapid exploration of chemical reaction space by performing multiple experiments in parallel, enabling the simultaneous exploration of various reaction parameters.[44-46] While HTS has a rich history in the field of biology, its adoption into chemistry is only a recent occurrence. Even then, its adoption has been mainly limited to an industrial setting with, for example, the acceleration of pharmaceutical processes.[35, 44]

To streamline this arduous process, various data science tools have recently emerged through the development of methods relating chemical structure to function.[46] A recently developed method by Sigman and co-workers utilizing multivariate linear regression models is capable of simultaneously identifying better performers during the early stages of optimization while also gaining valuable insight into the mechanistic aspects of a reaction.[8, 47] The models are based on a mathematical relationship between the reaction outcome and both experimental and computational molecular descriptors.[8] These descriptors can then be incorporated into various data science workflows, with the goal of optimizing a specific outcome of a reaction, such as yield or regioselectivity. However, it should be noted that high performance in one objective does not necessarily correspond with the desired performance in another objective.[46] Many ML algorithms and other statistical methods can be employed to explore the multi-dimensionality of reaction optimization. For instance, algorithms may be used for classification tools for the exploration of reactivity cliffs, linearly regressing the structure to function, and reducing

dimensionality for the mapping of different reactivity patterns.[46] Linear models are by far the most common, even though other methods are just as applicable and effective for quantitative analysis yet less applied in the analysis of catalytic systems.[8] A general workflow for developing models is presented in **Figure 10**.

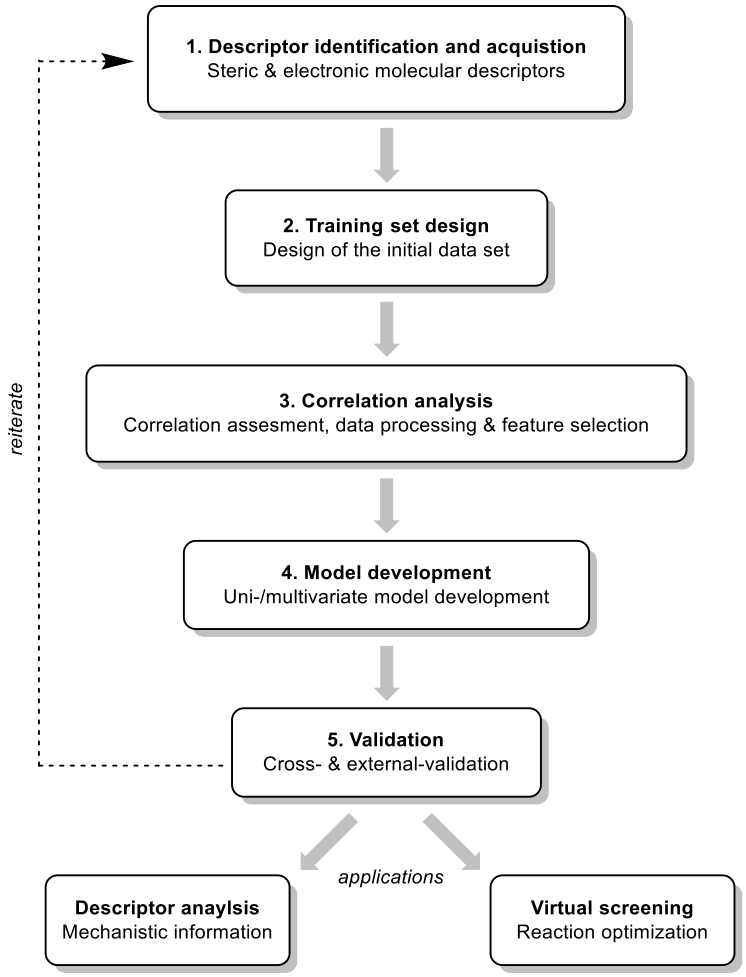


Figure 10. The general workflow for reaction model development.[8]

The general workflow consists of a five-step process: identification and acquisition of relevant descriptors, the design of the initial data set for model construction, intercorrelation assessment and feature selection, preliminary model development, and finally, the cross- and external validation of the developed model. Depending on the results of the last step, one may reiterate the process in order to improve the model or employ the model for real-world applications, such as the acceleration of reaction optimization via virtual screening or analysing descriptors for obtaining mechanistic insights.[8]

In order to attain accurate statistical models, a large library of descriptive features need to be built and acquired first.[8] For this, one may use large libraries of pre-existing data compiled from HTS or separate data sets from the literature. Although the organic literature might contain a large volume of data, it is often stored in various unstructured formats.[35] A more modern and structured database for the physical organic descriptors of phosphine ligands, KRAKEN, was developed recently by the Sigman group in 2022.[48] Even then, these libraries might not always contain the less-explored and desired molecules or reactions. Consequently, the generation of accurate molecular structures is still necessary. Methods such as DFT and/or the semiempirical extended tight binding (xTB)[49-52] based methods can be utilized for this purpose. Comprehensive analyses of xTB and DFT are outside the scope of this thesis and as such further discussion of these methods will be limited to a general overview.

As mentioned earlier, chemical reactivity can be roughly divided into steric and electronic effects. Molecular descriptors are, in essence, experimentally-derived or mathematical representations of these properties and effects.[35, 53] Furthermore, descriptors can be divided into local or global parameters, with global descriptors representing the molecule as a whole and local ones describing the properties of certain parts of the molecule.[54] Although individual descriptors may work well for the analysis of experimental data where a single effect dominates, such circumstances are rare in organometallic catalysis. A wider range of descriptors must be taken into consideration to encompass the complexity of both steric and electronic effects.[55] A comprehensive understanding of what various descriptors signify is essential for accurate model building and the acquisition of mechanistic insights.

Steric effects have been shown to play a vital role in influencing the regio- and stereoselectivity's of various reactions, such as cross-couplings[2, 6, 22] and asymmetric induction.[8, 55] This has resulted in the creation of quantitatively descriptive parameters for describing the steric effects imparted by ligands. Various steric descriptors have been previously introduced in literature, such as the earlier mentioned Taft parameter or the Tolman cone angle.[8]

The Tolman cone angle was originally introduced as a steric metric for phosphine ligands based on space-filling models of different ligands in $Ni(CO)_3L$ complexes, with a Ni—P distance of 2.28 Å. The angle is measured by projecting an arbitrary cylindrical cone from the central metal atom towards the outermost atoms of the ligand. Tolman's parameters have been successful in helping to understand various reaction outcomes.[7] However, the Tolman cone angle's contemporary use has become limited due to its inability to describe more structurally complex and asymmetrical ligands, such as bidentates, Buchwald-type biarylphosphines and NHCs.[8] Thus, percent buried volume ($\%V_{bur}$) was developed as a steric parameter to better represent the steric bulk of NHCs. It is defined as the percentage of the volume of an abstract metal-centered sphere (r = 3.5 Å) that is occupied by the ligand.[55]

Other commonly used steric parameters include bond lengths, bond angles, sterimol values[56], and bite angles. As with most steric parameters, atomic coordinates from both crystallography and calculated geometry-optimized structures can be used. Bite angles in particular are easy to measure from atomic coordinates. The descriptor is comparable to the Tolman cone angle, although for bidentate ligands in lieu of monodentates. The bite angle likely measures the net interaction with the metal center of the complex, capturing both steric and electronic effects.[8, 55]

While steric descriptors encompass local features such as bond lengths and bond angles[57], it is with electronic descriptors that the distinction between local and global parameters becomes especially apparent. Local electronic descriptors comprise atomic and natural bond orbital (NBO) charges, infrared (IR) frequencies, and site-specific nucleophilicities and electrophilicities. On the other hand, global descriptors include HOMO- and LUMO-energies, ionization potential (IP), electron affinity (EA) and dipole moment, which can be derived from the molecular electrostatic potential (MESP).[8, 54, 57] In addition to only representing local or global electronic characteristics of a molecule, electronic descriptors also contain valuable structural information. As such, manipulating the electronic properties of ligands proves to be an advantageous tool in the design of new catalysts.[8] Some of the aforementioned steric and electronic parameters are visualized in **Figure 11**.

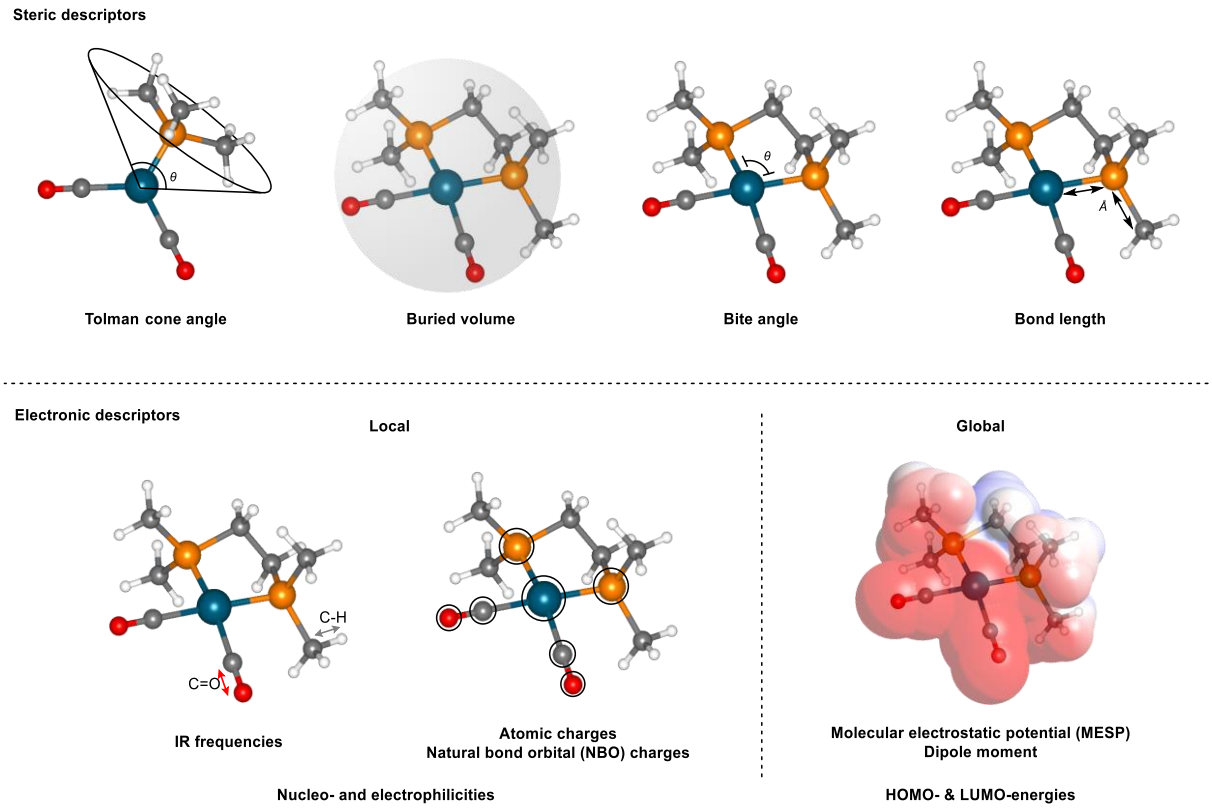


Figure 11. A selection of commonly used steric and electronic descriptors in the parametrization of transition metal complexes. Electronic parameters are divided into local and global descriptors.

Ultimately, the chosen descriptors should aim to capture both the steric and electronic characteristics of the molecules, while simultaneously conveying valuable mechanistic information.[57] Although prior mechanistic knowledge can aid in the selection of meaningful descriptors, it is also common to utilize various data science techniques, such as intercorrelation analysis or principal component analysis (PCA). These methods can help simplify and improve the model interpretation by reducing the dimensionality of the data and avoiding the usage of irrelevant parameters. Before any of the above-mentioned methods are put into practice, the data should be first and foremost normalized. Parameter normalization allows for comparison of model coefficients by normalizing all of the parameters to have the same scale and deviation.[8]

Descriptors that have significant intercorrelations with each other should be avoided in multivariate correlations. The presence of collinear parameters in a model causes the

effective variance to become associated with the difference between parameters, amplifying the random noise in the data. An additional benefit of intercorrelation analysis is that it helps benchmark descriptors whose physical meaning might be unclear against well-defined experimentally-derived descriptors. PCA is especially useful for models that are extrapolation-oriented. In essence, the method creates a new set of orthogonal parameters by maximizing the variance in the original data set.[58] PCA is extensively applied in dimensionality reduction, which, in addition to improving the modelling efficiency, reduces the concern of collinearity. However, PCA may also diminish the mechanistic information available in a data set, as the new reconstructed parameters tend to be less meaningful and harder to interpret.[8]

When modelling novel and less-explored reactions, the number of data points available can occasionally be quite limited. Consequently, the modelling outcome relies heavily on the careful selection of the initial data set used for model construction as to prevent the emergence of potential biases in the model. A biased model is limited to explaining the data at hand and lacks the ability to predict anything meaningful beyond the initial data set. To avoid these biases, the experimental data is typically divided into training and validation sets. Ensuring that the training set covers a broad enough spectrum of chemical reaction space and entries with lower performance is essential for building models that can effectively generalize to both the training and unseen data. However, having too much variance in the data poses the risk of overfitting. Overfitting occurs when the model attempts to explain all the random noise, making it overly specific towards the training set with poor generalizability to unseen data.[8] Indeed, finding the right balance between bias and variance is what makes training set design so challenging, as having both is typically impossible.

Due to the complex nature of organic reactions, chemists have long sought to simplify the overall problem. As stated earlier, a common approach in the past has been the use of linear models based on free energy relationships, or LFERs. In this approach, the relationships are identified and then solved using uni- or multivariate linear regression (MLR) models. These models have allowed for the extraction of valuable information about the underlying mechanisms, while needing a comparatively small number of data points.[35, 39]

An alternative and a more general approach to the problem of reactivity prediction is the utilization of ML algorithms. These statistical methods had not seen widespread use in organic chemistry until fairly recently, primarily due to the lack of reaction data in the past. Recent advancements in methodologies for reaction data generation and compound libraries have opened the field of reactivity prediction to ML different models, such as random forests, support vector machines and artificial neural networks.[7, 35, 39] Models capable of classifying non-linearly separable data are of particular interest. The functionalities of some of the above-mentioned ML methods are illustrated below in **Figure 12**.

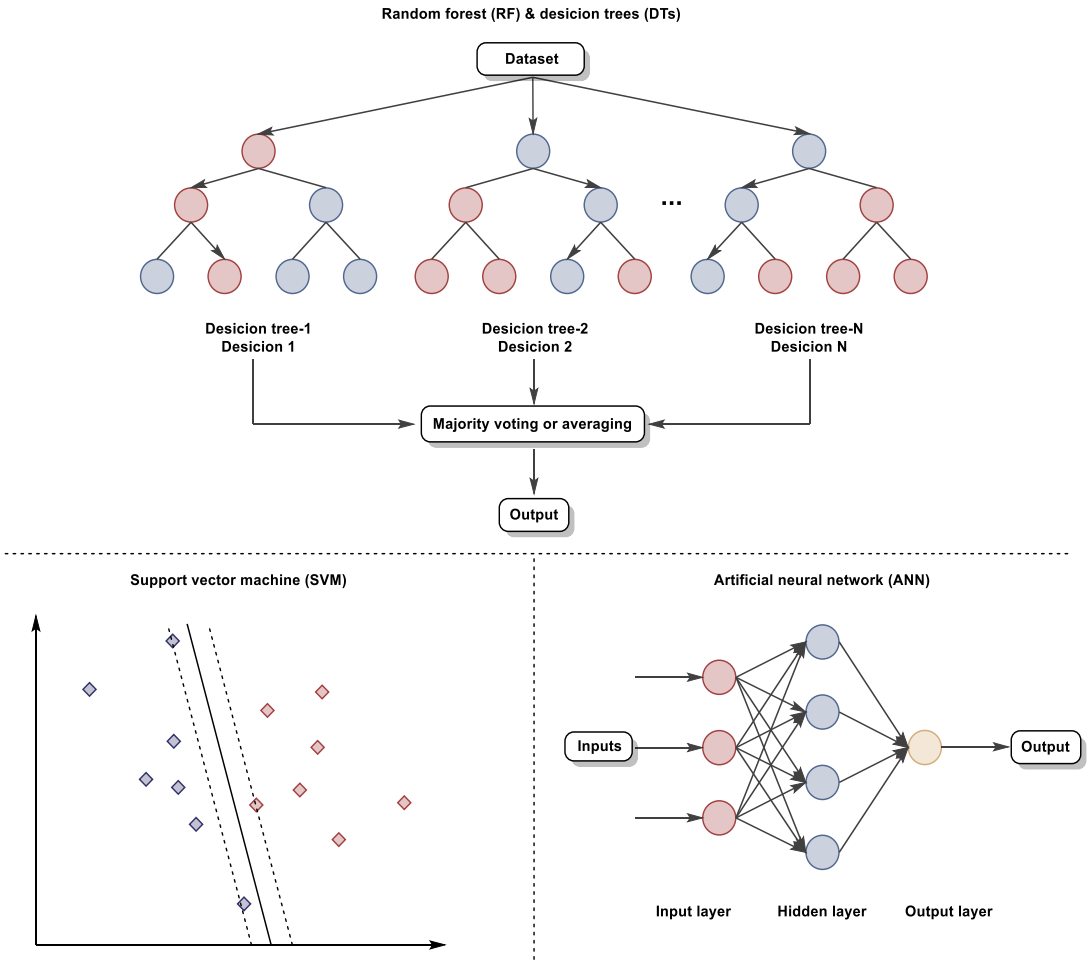


Figure 12. Illustrations of different ML models. Random forests utilize multiple decision trees for the output. Support vector machine divides the points in space by hyperplanes. Artificial neural networks comprise of a series of artificial neurons that receive, process, and output data.[39, 59]

Decision tree (DT) classifiers utilize a recursively built tree-like model of simple decisions, based on the data features. Each node of the tree either represents a numerical value or a class label, making DTs able to hand both numerical and categorical data. In addition to being used for classification purposes, DTs may be used for regression analysis. A deeper tree generally signifies more complex decision rules and thus a fitter model. However, DTs are at risk of creating overly complex trees that do not generalize to the data well, that is, overfitting.[59]

To reduce overfitting, an ensemble of randomized DTs are employed in a random forest (RF) classifier. Several independently constructed trees are fit in parallel on random subsets of the data. Each tree then votes on the predicted class and their outputs are either averaged or majority voting is used for the final result. This often leads to more accurate and robust predictions than when using individual trees. In addition, RFs can be used to measure feature importance based on the contributions of each features to the accuracy of the model. Other extensions of DTs include extreme gradient boosting (XGBoost) and adaptive boosting (AdaBoost), an ensemble learning algorithm that creates a stronger learner by combining predictions from weak learners, such as shallow DTs.[59, 60]

Support vector machines (SVMs) map data points in higher- or infinite-dimensional space that are then used to construct a hyperplane or a set of hyperplanes. SVMs search for the most optimal hyperplane which has the greatest distance, or margin, from the nearest data points of any class. The greater the margin, the lower the classifier's generalization error.[59] The behaviour of the separation is based on different mathematical functions known as the kernel function. The selection of the kernel function is generally based on the nature of the data, that is its dimensionality and separability. Different functions include linear, polynomial, gaussian, sigmoid, and radial basis function (RBF). While SVMs are generally used for classification purposes, they can be also used for regression analysis.[61] SVMs have found usage in chemistry since the early 2000s in QSAR and protein structure studies.[35]

Artificial neural networks (NNs) are complex structures consisting of interconnected nodes that are combined into several processing layers. These layers typically consist of

an input layer, one or more hidden layers, and an output layer. The nodes in one layer are connected to each node in the following layer. Each connection has an associated weight that represents the strength of the connection. A mathematical activation function then determines the output based on the weighted sum of its inputs. NNs with multiple hidden layers are commonly referred to as deep learning algorithms. Common deep learning algorithms include convolutional NNs and multi-layer perceptron's (MLPs). The weights of the hidden layers are adjusted when an algorithm is learning from the training data. MLP in particular utilizes a backpropagation algorithm, an iterative process where the weights are updated based on the error between predicted and actual outputs. NNs like any other ML models are susceptible to overfitting, generally as a result of too many nodes in the network.[62] The interest in NNs in the field of chemistry has grown rapidly since 1986 and has been investigated in QSAR studies since the 1990s.[35]

The most common method for evaluating the generalizability of a model is to use both cross- and external validation sets. Cross-validation is performed on the training set by excluding a part of the original data and constructing a different model with the remaining set of the training data. The excluded data is then predicted using the new model. Prediction accuracy of the excluded data gives an indication of the model's stability and generalizability. One of the several approaches to cross-validation is leave-one-out analysis, where each data point is excluded out of the set and tested individually. In other approaches it is common to average the outcomes across multiple runs. In comparison, in external validation, a separate set of data outside the training set is used to verify the model's accuracy. In this set, the empirical results of the set are known before model development.[8, 63]

Lastly, the applications. As briefly stated earlier, these models may be employed for different real-world applications. One such application is the virtual screening of lead compounds in drug discovery. Although HTS has been used to expedite pharmaceutical processes in the past, it is still a time-consuming process that involves conducting assays on thousands of different compounds. Rather than physically testing every compound in a laboratory, virtual screening entails predicting the reaction outcomes *a priori*. This can significantly narrow down the number of compounds necessary for laboratory testing, saving time and resources.[64, 65] However, it should be noted that the evaluated structures

should lie within the generalizable region of the models. Structures differing too much from the training set can lead to prediction failure. Averaging the predictions from multiple models can help improve the accuracy of the estimations. Other important factors, such as the commercial availability of the starting materials and the practicality of synthesis, should be accounted for when designing the virtual screening deck.[8]

Another application is the analysis of relevant descriptors for obtaining mechanistic insights. It should be stated that this type of analysis cannot be performed successfully if a prior hypothesis for the reaction mechanism does not already exist. This is indeed a powerful tool for obtaining mechanistic rationales. Parameters of discernible physical significance or correspondence with existing mechanistic information are generally selected for further analysis. Descriptor analysis may also be employed alongside with virtual screening techniques and guide the screening process towards a more focused library of simulated structures.[8]

When designing mechanistically informative models, simplicity and interpretability are key. Models comprising of a small number of parameters with clear physical meaning are often preferred over complicated models. Strong intercorrelation among the parameters should also be avoided, as the resulting features are defined by the differences between the parameters rather than the individual features described by them. This makes them generally difficult to interpret. Due to the unavoidable interrelationship between parameters, multiple models with different parameters can be utilized instead. Where obscurity and complexity would be vital flaws for mechanistically-driven studies, they are typically inconsequential in purely explorative modelling. In the latter case, accuracy and generalizability take precedence.[8]

2.2.3 Machine learning in cross-coupling reactions

Although the use of ML in chemistry dates all the way back to the 1990s, its application to cross-coupling reactions is a fairly recent development. Various factors have previously limited the use of ML in problems related to chemical synthesis. Namely, the implementation of these algorithms has historically been limited to specialists in that field only. Furthermore, the scarcity of reaction data made the obtainment of statistically

meaningful results extremely challenging. ML applications have finally become feasible in this field due to the recent advancements in HTS and the increased accessibility of model development.[39, 66] While some of the more general applications of ML in this field were briefly covered in an earlier chapter, this chapter aims to highlight specific examples of how ML has been applied in various cross-coupling reactions.

Many of the previous endeavours have focused on employing ML to aid in synthesis planning and for predicting reaction outcomes. However, predicting the performance of a given reaction remains a much rarer occurrence. With this in mind, in 2018, Ahneman and colleagues from Doyle's group conducted a comprehensive study into predicting the performance of various isoxazole additives in palladium-catalysed Buchwald–Hartwig cross-couplings. As isoxazoles have drug-like characteristics but are underrepresented in successful drug candidates, the team sought to explore the effects of isoxazole fragments on the amination of different aryl and heteroaryl halides. Glorius fragment screening[67] was conducted to generate data instead of directly evaluating couplings of substrates containing the heterocycle functionality. Using a synthesis robot, the team performed a total of 4,608 reactions using 23 different isoxazoles. A set of 120 molecular, atomic, and vibrational descriptors were calculated. The descriptors were then used as inputs and reaction yield as outputs for the predictive models. Rather than modelling the performance of catalytic systems, the study focused more on comparing the performances of various linear regression and machine learning methods. Their developed RF model was successfully applied to sparse training sets and out-of-sample prediction.[66]

Finally, the authors sought to determine the feasibility of employing their RF model in guiding mechanistic analysis. Direct interpretation of RF models presents a considerably greater challenge compared to that of linear regression models. Where a direct comparison of the contribution of each feature to the model's predictions can be made in linear regression models, RFs require the evaluation of the relative importance of the descriptors employed in constructing the model. Their conducted analysis of these descriptors suggested that the propensity of the additive to act as an electrophile influenced the reaction outcomes. Furthermore, they hypothesized that competitive oxidative addition of the isoxazole could contribute to the unfavourable side reactivity.

While oxidative addition of Pd to isoxazoles is not known, such elementary steps have been reported for other transition metals in previous studies.[66]

Continuing with theme of Buchwald–Hartwig couplings, in 2023, Rinehart and co-workers developed a ML tool for predicting substrate-adaptive reaction conditions. The modelling of Buchwald–Hartwig couplings is a multi-dimensional problem, which includes factors such as the utilized nucleophiles, electrophiles, catalysts, solvents, and bases. Each individual dimension can impact the yield, and each reactant may also exhibit different preferences with regard to catalysts, solvents, and bases. As existing databases lack the complexity required to effectively model these relationships, the design and construction of their model required the generation of an entirely new data set. Their initial reaction space comprised of 180 conditions (3 bases, 3 solvents, and 20 ligands) and 2,500 substrate pairs (50 amines and 50 bromides). This added up to an experimental impossibility of 450,000 possible reactions. However, with clever exploration of the reaction space and pairings, they were able to reduce the experimental burden down to 3,300 experiments.[68]

Their final data set included over 120 reactant pairs, which allowed models to learn nonlinear reactivity trends across various reactant classes. For the model itself, Rinehart and colleagues used deep feed-forward NNs and an active-learning workflow for the training of the models. Validation of the model was conducted with couplings in which one or both reactants were not seen by the model and were tested experimentally. For reactants that were well-represented in the reaction space, predictions were generally robust. Structural permutations were correctly predicted with learned activity trends. However, the performance of the prediction ranged from moderate to poor for new types of structures that had differing reactivity patterns from those present in the data set. Expanding the dataset's domain improved poor model performance, demonstrating the workflow's ability to continually improve prediction capability with the accumulation of data. Overall, their model demonstrated good generalization to out-of-sample substrates and a mean absolute error of 9% in predicting reaction yields using randomly partitioned data.[68]

In 2022, Angello and co-workers sought to find more general reaction conditions for heteroaryl SMCC's with the assistance of ML. Efforts to identify such general conditions often require exhaustive experimentation and the consideration of vast regions within the realm of chemical space. This search space encompasses all possible combinations of substrates and reaction conditions, rendering navigation through standard approaches impractical. Thus, finding general conditions for SMCC's remains as a challenging and largely unsolved problem. Their approach to solving this problem was a closed-loop workflow, that leveraged data-guided matrix down-selection, uncertainty-minimizing ML, and robotic experimentation.[69]

To facilitate the practical exploration of general SMMC reaction conditions, they first streamlined the selection of both the possible substrates and reaction conditions. This strategic down-selection process aimed to preserve the relevance of the subsets to their wholes. After streamlining the reaction space, they performed a total of 528 reactions on a robotic system. According to the authors, this de novo approach was chosen with the aim of generating small sets of highly reproducible data during closed-loop optimization. The data would also include negative examples, which are vastly underrepresented in literature. A strategic focus was also placed on decreasing the model's uncertainty and maximizing the efficiency of the learning process. Applying this workflow to SMCC's allowed them to identify conditions that doubled the average yield relative to a widely used benchmark that was developed previously using traditional techniques.[69]

Another application of ML to SMCC's was conducted by Meyer *et al.* back in 2018. In this study, the authors developed a ML model that can predict the energy of the oxidative addition process between a transition metal complex and a substrate. This could be then used as a descriptor to estimate the activity of homogenous catalysts using volcano plots. Volcano plots function by utilizing Sabatier's principle, that is, an ideal catalyst should not bind a substrate too strongly or too weakly. Species positioned near the peak of the plot exhibit the most favourable thermodynamic profiles, and thus, adhere to Sabatier's principle. Conversely, those found along the right or left slopes possess less ideal profiles and have either overly weak (right) or overly strong (left) substrate/catalyst interactions. This simplified thermodynamic approach was used to swiftly discriminate between catalysts with promising or inadequate energy profiles.[70]

The authors assembled a database comprising of 25,116 transition metal complexes. In total, 7,054 of these complexes and their computed binding energies would be used for the training of the ML model. The remaining 18,062 homogeneous catalysts were virtually screened using their molecular volcano plot model to identify 557 thermodynamically attractive candidates. The number was further narrowed down 37 by searching for low-cost complexes. The finalists mainly consisted of palladium and copper complexes with both standard and less common ligands.[70]

With large language models (LLMs), such as OpenAI's ChatGPT, having recently emerged as powerful tools capable of interpreting and answering questions via verbal human input, it is unsurprising that it has already been utilized in chemical synthesis-related problems. Although ChatGPT is not directly a model for chemistry, its training data includes vast amounts of scientific literature. Taking this into account, Mahjour, Hoffstadt, and Cernak investigated the matter of utilizing ChatGPT for the generation of HTS reaction arrays for common cross-coupling reactions, such as Suzuki–Miyaura and Buchwald–Hartwig. The authors began their study by tasking ChatGPT with generation of experimental designs for various pairs of substrates for each different reaction class. While ChatGPT was able to successfully formulate these reaction arrays, the authors doubt that the model exercised any physical or chemical intuition in its designs. The model appeared to select popular reagents and catalysts associated with the reaction-type keywords. However, it may be less aware of how they are actually used together. Nonetheless, this resulted in interesting and feasible proposals for array recipes, with the desired product observed in every instance tried on the first attempt, and isolated yields ranging from 62 to 94%.[71]

While there have been major successes in this field, it is not a rare occurrence that no satisfactory model can be found. In 2022, Beker and co-workers attempted, but failed, to use ML to discover general reaction conditions for SMCC's through extensive mining of chemical literature. ML applications of chemistry generally operate under the assumption that a large volume of literature examples should enable the creation of precise and predictive models of chemical reactivity. This is usually the case when the descriptors used to construct the model can adequately capture the chemical essence of the problem. Even though the authors constructed their database from over 10,000

carefully curated examples, ML methods failed to demonstrate any notable improvement over naïve assignments based solely on the frequency of certain reaction conditions reported in literature. This result held regardless of the employed model, or the representation used to describe the reaction partners. The authors mainly attributed this failure to the lack of negative examples and causal links to the substrate's structures in current literature. Their suggestion to solving this problem is to begin augmenting existing literature data by standardized and systematic experiments. In these experiments, reactions would be repeated under multiple different conditions, which should give more meaningful conclusions about better and worse reaction conditions.[72]

3 Experimental section

The experimental part of this thesis was carried out at the Orion Corporation in the department of Medicinal Chemistry from January 2023 to June 2023. Similarly to the literature review, this section has also been divided into two main categories: computational modelling and laboratory testing. A large part of the computational study and model development was performed on my personal computer (Windows 10 & WSL Ubuntu, AMD Ryzen 7 2700X, NVIDIA GeForce GTX 1070 & 16 GB of RAM) in addition to the laptop provided by the Orion Corporation (Windows 10, Intel i5-10310U, Intel UHD Graphics, 16 GB of RAM). DSW cloud computing was employed for the geometry optimizations and generation of conformer ensembles. Descriptor calculations were performed on my personal computer with Ubuntu for Windows.

With the recent and rapid development of AI, its significance and utilization has increased dramatically in our everyday lives. Similar developments have been witnessed with ML in the field of predictive modelling and optimization of the medicinal chemistry processes in drug discovery. Although the application of AI and ML are still at the very early stages, their potential to revolutionize the drug discovery process has already been demonstrated.[10] In the near future, ML models will serve as complementary methods with the aim of guiding chemists in complex reactions by pointing them in the right direction. By suggesting a wide range of reaction conditions for the attainment of a specific outcome, the need for extensive guesswork should diminish considerably. Contrary to methodology-oriented studies, minimizing the time spent on unsuccessful experiments is essential for drug development. Failed reactions serve only to impede progress and interrupt workflows, further delaying the development of new potentially life-saving drugs.[65, 73, 74]

While numerous tools have been developed and are routinely utilized, there are still many aspects of the drug discovery process that can benefit from further refinement of these computational methods and tools.[10] As such, the primary goal of this thesis was the development of a predictive model for ligands affecting the regioselectivity's of various *N*-heteroarenes and the tools and workflows accompanying it. The developed model should serve more or less as a proof of concept while simultaneously laying the

groundwork for future model development with relative ease. These models and tools can be utilized to model a large variety of commonly used reactions in medicinal or other fields of chemistry, such as the SMCC reaction. Although this thesis specifically focuses on palladium catalysed SMCC's, all the developed tools and workflows were developed with a focus on future expandability and adaptability. Accommodating other substrate classes or even entirely different reaction types is easily within the realm of possibility. The end goal of this thesis would be the integration of predictive models into laboratory notebooks or other comparable software used for reaction design, but such objectives remain out of reach with the given timeframe. The general workflow utilized for the model development in this thesis can be roughly divided into five distinct steps. It is presented below in **Figure 13**.

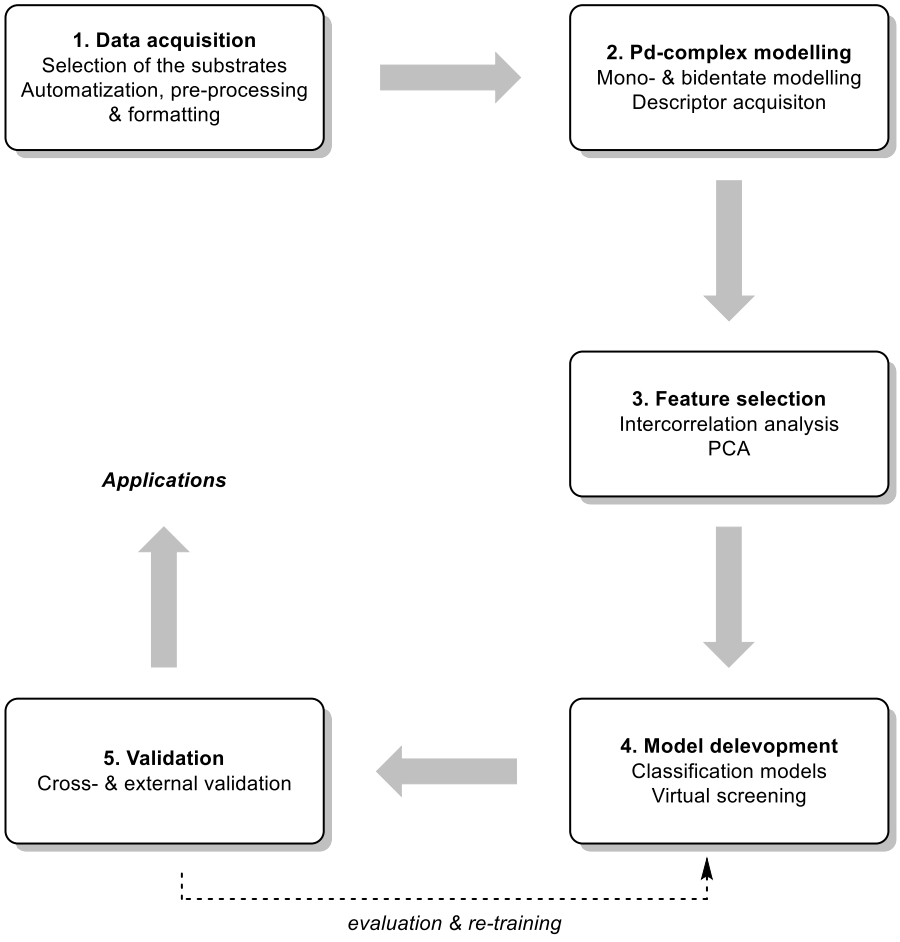


Figure 13. The general workflow of the model development in this thesis.

The study began with selection of the substrates and the corresponding SMCC reaction data. Various deciding factors, such as substrate properties and the amount of data

available, guided the selection process. As this data had to be acquired from an external database, making it ML friendly required multiple layers of automatization and pre-processing. After rigorous formatting, ligands used in the palladium complexes of the reactions were identified and utilized for the generation of the initial transition metal complexes. The generated complexes geometries were optimized and used to generate the conformer ensembles. These were then subsequently parametrized. Intercorrelation analysis and feature selection processes was performed on the acquired descriptors to find the most relevant features, which were then used for the training of the classification models. Various models and their performances were first examined, and the most promising models were used for the virtual screening of various commercially available ligands. Finally, the best ligands were tested in laboratory and the experimentally obtained data was used to validate and re-train the model.

3.1 Computational modelling

3.1.1 Substrate selection

The work began with the first step in model development: determining what type of data was necessary to build the model. It was established very early on that the model in question would be used for predicting the site-selectivity's of various ligands for a given *N*-heteroarene substrate. In other words, a ligand classification model for catalyst-controlled approaches. After a brief literature review, it was decided that palladium catalysed SMCC's of *N*-heteroarenes would be the modelled reactions in question.

A lot of thought would have to be put into the selection of the heteroarene substrates. Both organic chemistry and data science perspectives would have to be accounted for when selecting the data. The regiochemical outcomes of *N*-heteroarene SMCC's are strongly influenced by a myriad of factors, including the intrinsic propensity of each halide to undergo oxidative addition.[22] To properly assess unconventional site-selectivity, the substrates would also have to bear identical halides, which considerably limits the amount of data available. As previously mentioned, one of the biggest limiting factor for the development of ML models in the past has been the lack of data.[39] Enough of verified data on very specific reactions with specific reactants would have to be

obtained. Additionally, this data should preferably be in a structure format suitable for ML applications. Other parameters, such as reaction conditions would also have to be included in the data, as a broad enough coverage of different conditions is essential for building a good model.

As luck would have it, Norman *et al.* had recently investigated the matter of *N*-heteroarene cross-couplings. Their study included a table of various dihaloheteroarenes with their reported occurrences of unconventional site-selectivity in literature.[1] Thus, these would be the substrates that would form the basis for the reaction data. The heteroarenes analysed by Norman and co-workers are presented in **Figure 14**.

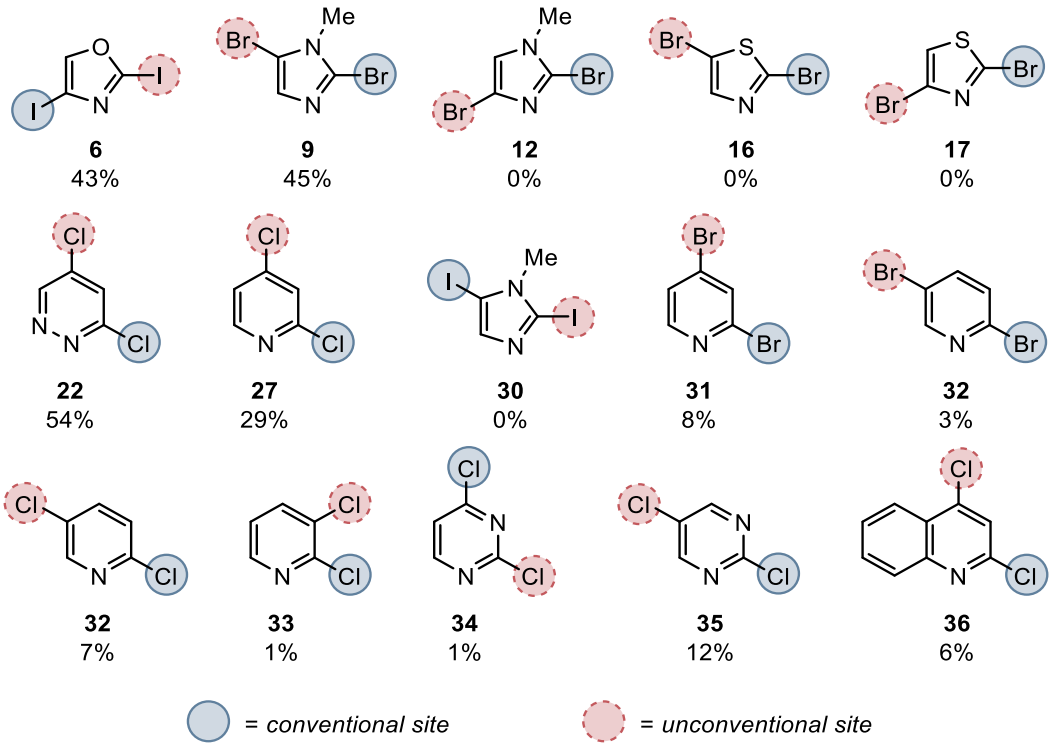


Figure 14. N-heteroarene substrates with the conventional and unconventional sites highlighted. The percentage values indicate the ratio of Pd-catalysed cross-coupling reactions favouring unconventional site-selectivity, according to a SciFinder search conducted by Norman *et al.*[1]

A similar SciFinder search was conducted to verify their results for the palladium catalysed SMCC's. Out of the fifteen substrates analysed, only a handful warranted further investigation. For instance, substrates whose number of overall examples or the

percentage of reactions favouring the unconventional site were too small to evaluate were omitted altogether. In addition, it was decided to limit the selection of the substrates only to six-membered heterocycles, primarily due to them being the most studied and best understood, thus making them the best candidates for building a predictive model. The inclusion of both five- and six-membered heterocycles in one model would probably not have been feasible in the first place. Apart from the obvious structural differences, the electronic characteristics of five-membered heteroarenes differ from their six-membered counterparts, making them less reactive electrophiles. The small number of publications on the topic of five-membered arene cross-couplings also complicates the definition of "conventional" selectivity for these substrate classes. Furthermore, the mechanistic origin of the ligand-controlled divergent selectivity with some five-membered heteroarenes still remains unknown.[1]

Unfortunately, none of the studied substrates fit the aforementioned criteria perfectly. Ideally, there would be a large amount of data available with a number of different reaction conditions. A large majority of the inspected reactions had nearly identical conditions, with DPPF and PPh_3 accounting for majority of the utilized ligands. However, two substrates in particular fit the criteria better than the rest: 3,5-dichloropyridazine (**22**) and 2,4-dichloropyridine (**27**). 3,5-dichloropyridazine had an ideal ratio of around 50/50 for reactions preferring conventional and unconventional site-selectivity, yet the small number of overall reactions (107 reactions) was less than ideal. 2,4-dichloropyridine on the other hand had a lot more reaction data available in literature (846 reactions), although the number of reactions displaying unconventional selectivity was far outnumbered by its conventional counterpart. Nevertheless, the amount of data in both cases was deemed sufficient enough to continue to the next step.

3.1.2 Data acquisition and pre-processing

After the decision of what the training data should roughly consist of, the next obvious step in model development was data acquisition. The data was to be extracted from external reaction databases like SciFinder and Reaxys. Although these databases contain a large volume of data, it too suffers from the very same problem that has ailed organic literature to this day. More often than not, the data is stored in different unstructured

formats and contains only the highly optimized reactions. Failures, contradictory outcomes, and side reactions are generally omitted from the final publications, data which is often crucial for the development of good ML models.

Nevertheless, the work began on data acquisition with SciFinder. These databases are generally designed with the end-user in mind, making them convenient to the chemist for finding singular reactions. However, for the purposes of procuring large volumes of uniform and well-structured data, these databases remain to be rather unsuitable. As such, the data was to be manually extracted and processed. Cataloguing hundreds if not thousands of different reactions by hand can prove to be quite a daunting task. A task which is typically better left to the computer. Initially, the reaction data was manually catalogued and written to a spreadsheet. Careful consideration was put into readability, future expandability, and computer processability. This way the data would be easily readable by both the chemist and a computer. However, questioning the feasibility of writing down everything manually, various programs for data acquisition and processing were written as a consequence. What initially started as simple Python scripts that would ease data extraction from a select part of the database quickly grew into a fully automated system capable of processing hundreds of reactions in a minute. A considerable speedup when compared to going through the same number of reactions by hand. The data acquisition process is visualized in **Figure 15**. The biggest downside of this system, however, was that the validity of the extracted data would have to be manually confirmed at a later stage.

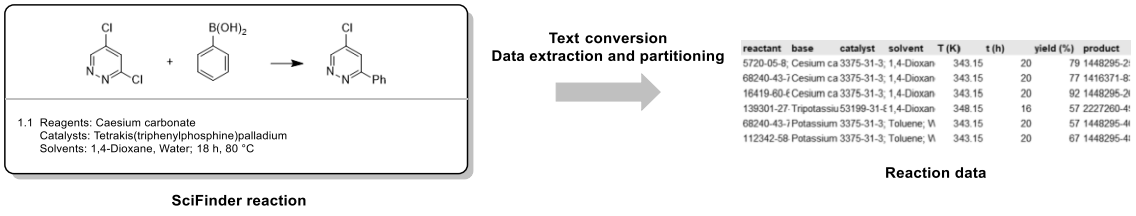


Figure 15. Reaction data from SciFinder is automatically extracted, converted, and partitioned to the appropriate spreadsheet using Python.

In essence, the program examines the pages of the database provided by the user for all its reactions. The desired information is then extracted from the reactions, with the acquired data being partitioned into different categories, each noted with a different

suffix. Various factors, such as the number of end products and the inclusion of certain essential information, affected this data partitioning. For the development of this model, features that were deemed essential were reactants, the catalyst, the base, solvents, reaction time, temperature, yield, and the end product. Various additional parameters, such as the chemical abstracts service (CAS) codes for all molecules, were extracted. As CAS codes provide a unique numeric identifier for a substance, a library containing various molecular properties or other useful parameters related to the molecule could be constructed. One such parameter was the simplified molecular input line entry specification (SMILES), which provides a 2D representation of a molecule in a string format.[39] Reactions that lacked the aforementioned details were subsequently omitted from the final data set. The final number of reactions in the data set for 2,4-dichloropyridine was 219, and 55 for 3,5-dichloropyridazine.

Although the program was capable of parsing through SciFinders various structured and structured formats, it was far from perfect. Errors could and would occur quite frequently. These were most commonly caused by the absence of some vital information, or the data was in an otherwise unsuitable or unstructured format. Since this type of data would require manual inspection later, these reactions were consequently saved on a separate spreadsheet. The obtained data required a lot for formatting before it could be utilized for ML or modelling purposes. This was especially apparent in cases where an otherwise one-pot reaction was divided into its component steps in the database. These reactions would be, if possible, automatically combined back to one step reactions. Considering the timeframe in which this was developed, the program proved to be more than adequate for the task it was designed for.

In retrospect, utilizing Reaxys instead of SciFinder would have likely proved to be an easier task. The data is often presented in a far more structured format than that of SciFinder's, but ultimately the use of Reaxys was omitted due to the nature of the website's design making data extraction far too tedious of a task. While this form of data acquisition is still the unfortunate reality today, it does not necessarily have to be so in the future. Standardizing the presentation of the reaction data in a more ML friendly format would be the ideal way going forward for these databases. This way less time

would be spent on data acquisition and formatting, leaving more time for actual model development.

While all the required reaction data had been gathered at this point, the data required further processing before any modelling or model development could commence. For instance, correct labelling of the data and analysis of the palladium complexes was still necessary. In ML, data labelling refers to the process of assigning informative and meaningful labels to raw data. When a model is trained on labelled training data, this process is known as supervised learning. Common types of supervised learning include regression and classification models.[59] The task of labelling the reaction data based on their outcomes was fairly easy to automate using RDKit[75], an open-source collection of cheminformatics and machine-learning software. First, however, a large proportion of the molecules had to be converted into their subsequent CAS codes and thereafter to the SMILES strings linked to the codes. While traditional naming conventions are useful to a chemist, they are largely an additional hindrance for a computer. By converting all the reagents and end-products to their SMILES strings, it was possible to fully automate the comparison process between the two and solve which site of the substrate had reacted. Conventional site-selectivity was denoted with zero, unconventional with one, and two was used if the end-product was diarylated. This process is visualized in **Figure 16**. It should be noted that while this thesis solely focuses on six-membered *N*-heteroarenes, the methods and workflows developed here should be applicable to any other six- or five-membered heteroarenes.

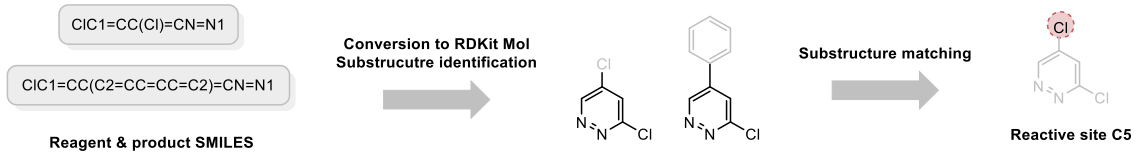


Figure 16. SMILES strings of both reagent and product are first converted to RDKit's own 2D molecular graph. The common substructure is then matched to the reagent to find the site of reactivity.

With the end goal being a ligand classification model for catalyst-controlled approaches, fragmenting the palladium complexes to their respective ligands was still a necessary step before continuing to the next step. Implementing this was also quite straightforward

with RDKit. By removing the metal center from the complex altogether, determining which atoms of the ligands had coordinated to the palladium became trivial. This information could be later used when creating the new metal complexes for modelling. Occasionally the ligands were already provided in the reaction data. As these were largely phosphines, it was assumed that phosphorus would be the most likely atom to coordinate with the metal center. Special consideration had to be given to ferrocene type ligands, such as QPhos. When converting from SMILES to RDKit's Mol structure, RDKit incorrectly assumed that the Fe^{2+} and the carbons of the cyclopentadienyl rings were all covalently bonded together. Altering the bond types from singular to unspecified appeared to largely correct this issue.

Further parameters could be easily derived from the fragmented complexes, such as the number of ligands coordinated to the palladium, that is ligation state, and the chelation of the individual ligands. Previous studies have highlighted that chelation may play a significant role in determining the regiochemical outcomes of SMCC's[4], which made these parameters particularly important. Determining the ligation state of the active catalyst was not something that could be easily ascertained from the SMILES string alone, and thus, this parameter would not see derivation through these means.

3.1.3 Palladium complex modelling and ligand parametrization

Before exclusively committing to the pursuit of parametrization through transition metal complex modelling, various alternatives were first examined. The simplest way to depict a chemical reaction is by a one-hot encoded vector, which indicates whether a specific compound is present within the reaction. While lacking chemical significance, such models can reveal statistical correlations in combinatorial data sets. Given that molecular descriptors inherently include chemical information, they may be employed to predict novel molecules outside of the initial training set. However, biased selection of descriptors run the risk of oversimplifying the problem, which may introduce systematic errors through the loss of seemingly unimportant information. Furthermore, the task of selecting consistent and universally applicable molecular descriptors is particularly challenging, due to the uniqueness of each molecule and reaction.[39]

As an alternative approach, the application of structural representations, such as molecular fingerprints, could be potentially used to circumvent the aforementioned issues. Developed by Sandfort and co-workers, their approach involves the generation of multiple fingerprint features (MFFs) solely from structural input, such as a SMILES string. Each molecule is converted to an array of 24 diversly configured fingerprints generated using RDKit. Molecular fingerprints are, in essence, bit vectors that contain information about the molecules various substructures which can be used to quantify molecular similarity using the Tanimoto coefficient.[39, 76] Multiple MFFs are matched to the observed experimental data and used to train a ML model. In the end, they successfully demonstrated the model's applicability to a range of diverse problems related to organic molecules.[39]

Their approach was of particular interest as it would have been relatively straightforward and fast to implement due their utilization of RDKit. Some preliminary work based on the implementations of Sandfort *et al.* was performed, namely the generation of the MFFs for all molecules and multi-hot tables for solvents and bases. The difference between one-hot and multi-hot tables is simple, as multi-hot tables can contain multiple entries instead of one. It was eventually decided that due to the three-dimensional nature and overall complexity of catalysts, a descriptor-based approach would be more fitting for the purposes of this thesis. However, the work performed here was not all for naught, as the multi-hot tables of bases and solvents were later adapted and converted to decimal values. The processes for both the fingerprint generations and multi-hot decimal encodings are presented in **Figure 17**. Although the structure-based approach did not see application this time, it could provide insights into the substrates or ligand structural effects on site-selectivity.

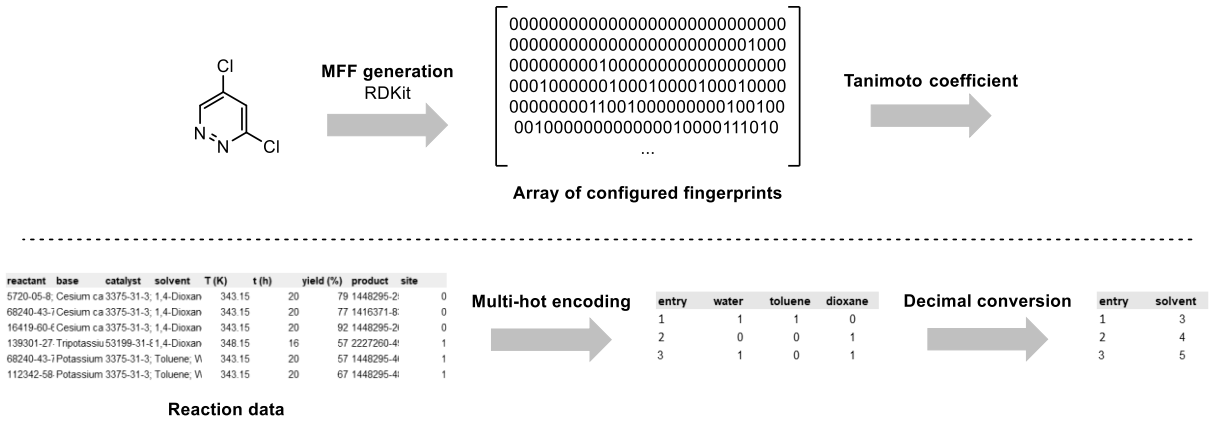


Figure 17. MFFs were generated for all molecules in the reaction data set. All utilized solvents and bases are presented as a multi-hot table, which were later converted to numeric values in a manner akin to decimal conversion of binary numbers.

With the fingerprint-based approach having been shelved at this point, the recently developed Sigman group's KRAKEN organophosphorus(III) descriptor library was instead selected as a starting point for the ligand parametrization. The library contains 190 conformationally representative descriptors for each of the several thousand monodentate phosphines. Because of the way ligands were parametrized, bidentate and nitrogen ligands are not included in the library.[48] Based on the findings of Lempinen, A., a previous thesis worker for Orion, KRAKEN descriptors alone were not able to sufficiently describe the reaction data of nickel catalysed SMCC's.[77] With the aim of avoiding unnecessary time spent on a potential dead-end, the decision was made to implement GFN-xTB and conformer-rotamer ensemble sampling tool (CREST)[51, 78] workflows for transition metal complex modelling akin to those previously designed in literature.[9, 48, 79, 80] The palladium complex generation and parametrization workflow developed for this thesis is presented in **Figure 18**.

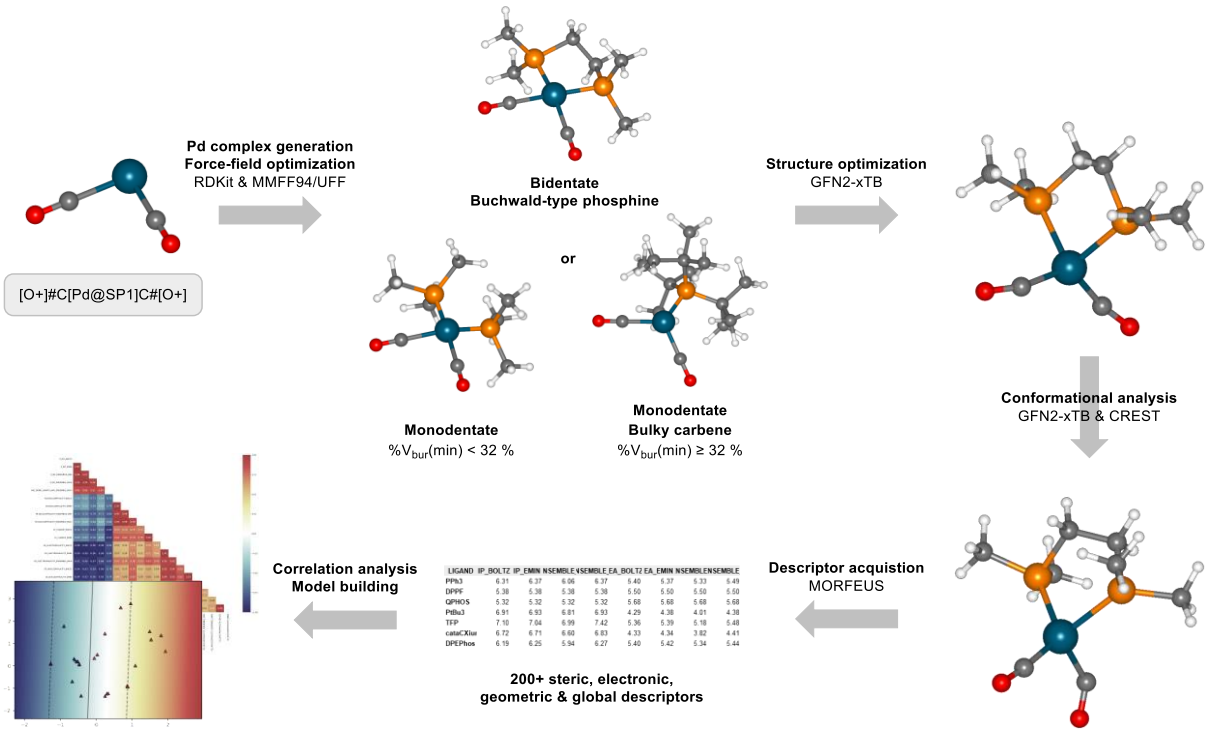


Figure 18. The general workflow for the palladium-carbonyl complex generation and parameterization using various ligands dependent on factors, such as denticity and $\%V_{bur}(min)$.

This implementation attempts to combine the modelling workflows of both mono- and bidentate ligands into one. While no common methods for both types of ligands currently exist in literature, the work carried out by Lempinen, A. would serve as the basis for the modelling methods implemented in this thesis. The nickel carbonyl complexes central in Lempinen's work are designed to model and minimize the transition state of reductive elimination. In this context, the carbonyls serve as analogues for the organic groups R^1 and R^2.[77] One of the distinguishing features of this work is that the initial geometries would be generated from $Pd(CO)_2$ fragments, instead of $Ni(CO)_2$. While palladium carbonyl compounds are not exactly stable[81], the decision to continue utilizing carbonyl ligands was largely influenced by the previous implementation's reliance of them, and my lack of prior knowledge and experience in transition metal complex modelling for the development of more novel workflows.

The metal-carbonyl bond is considered a σ-bond originating from the lone pair on carbon to an unoccupied orbital on the metal. It is reinforced by π back-bonding from a metal

orbital with suitable symmetry to an antibonding orbital of carbonyl. Factors such as the contribution of back-bonding and the overall charge of the complex can influence bond frequencies, distances, angles, and other steric and electronic properties.[82, 83] Thus, the steric effects of additional ligands would affect the carbonyl ligands, allowing for the calculation of descriptors from the metal atom and carbonyl ligands for both mono- and bidentate ligands. Additionally, the removal of carbonyl ligands should not be necessary before calculating the steric descriptors as they contribute equally to these values.[77] It should be noted that analysis performed on carbonyl ligands in transition metal complexes by Nandy, A. *et al*. indicated the average C-O bond length to be less than 1.128 Å when coordinated to palladium. Below this value, the σ- and ionic-contributions tend to dominate more than the contributions from π-bonding.[76] As such, the results discussed in the following chapters should be regarded with reservations in mind.

Incorporating both mono- and bidentate ligands required the investigation of the attainable ligation states of various ligands. While singular bidentate and Buchwald-type phosphine ligands were anticipated to bind to palladium in the requisite state for catalytic activity, monodentate ligands may attain both mono- and bisligated states. Based on the crystallographic studies of Newman–Stonebraker and co-workers, the determining factor between the two states appears to be the steric bulk of the ligand, or the KRAKEN database descriptor $\%V_{bur}$(min). For monodentate phosphines, values under 32% pointed to bisligated states and conversely, values over 32% to monoligated states. Values between 29% and 32% signified a region where bisligated states were thermodynamically favoured in the resting state, but monoligated states were still attainable. The equilibrium between the two states is influenced by the temperature and concentration of the reaction.[84] Inspired by this, the parametrization of the ligands would be performed based on their ligation states in the active catalyst and their $\%V_{bur}$(min) values. In addition, different ligand types would also affect their ligation states. The ligands examined later in this thesis and the parameters affecting their modelling are presented in **Table 1**.

Table 1. Denticities, ligation states and $\%V_{bur}$(min) values of the ligands used in this thesis. The percentage values were acquired by dividing the V_{bur}(min) values by the volume of the sphere or roughly 1.8 Å^3.

ligand	denticity	ligation state	$\%V_{bur}$(min)
PPh_3	mono	bis	28.2
P*t*-Bu_3	mono	mono	36.3
DPPF	bi	mono	-
QPHOS	mono	mono	41.0
cataCXium A	mono	mono	36.0
DPEPhos	bi	mono	-
t-BuDavePhos	mono	mono	36.6
TFP	mono	bis	25.6
P(*o*-tol)$_3$	mono	mono	34.3
PMe_3	mono	bis	22.1
$P(OPh)_3$	mono	bis	24.1
PCy_3	mono	bis	30.1
DPPPE	bi	mono	-
DPPB	bi	mono	-
DPPP	bi	mono	-
BINAP	bi	mono	-
Xantphos	bi	mono	-
Trimesitylphosphine	mono	mono	48.1
RuPhos	mono	mono	30.8
DavePhos	mono	mono	31.2
RockPhos	mono	mono	43.8
SPhos	mono	mono	31.6
XPhos	mono	mono	31.4

The generation of the initial palladium complex geometries was conducted using RDKit. Ligands from the reaction data were added to the $Pd(CO)_2$ fragments in accordance with

their ligation state, and the generated coordination geometry could be defined in the palladium fragments SMILES string. Bidentates and Buchwald-type phosphines were generated as complexes with square planar geometry. In contrast, monodentate ligands were generated as monoligated trigonal planar complexes if their $\%V_{bur}$(min) values were over the 32% threshold established by Newman–Stonebraker and co-workers.[84] Conversely, minimum buried volume values under 32% resulted in the generation of bisligated square planar complexes. Bulkier carbenes (primarily NHC's) were anticipated to form monoligated complexes and optimized as such. After transforming the complexes into three-dimensions, they were optimized with force field optimizations using either MMFF94 or UFF (**Figure 18**). However, general force-field methods usually lack the precision required for accurate optimization of transition metal complexes.[76] Alternatively, methods such as DFT or GFN-xTB may be employed instead and will yield more accurately optimized geometries.

Given that steric properties may vary significantly with conformation, the conformational space of each ligand has to be first analysed.[48] This is especially important, as previous studies have highlighted the significance of ligand sterics affecting the regioselective outcomes of SMCC's.[1] Although no individual model can comprehensively represent the entire conformational space available to a ligand in any given complex, defined boundaries for attainable geometries and properties do exist. Exploration of these boundaries can provide insights into the behaviour of the ligands in catalytic systems and predict their performance. The correct boundaries, however, can only be derived from a sufficiently complete ensemble of conformers. While DFT optimizations are considered fast for singular compounds, calculating hundreds of complexes can prove to be very computationally demanding. As such, GFN-xTB coupled together with CREST was utilized to generate the conformer ensembles at the fraction of the cost of DFT, albeit with a slight compromise in the accuracy of the generated geometries.[48]

The structural optimization of the CREST conformer ensembles was performed with GFN2-xTB level theory. Frequency calculations were performed using Hessian calculations, which assesses the gradient norms for non-zero values and eliminates imaginary vibrational frequencies. Descriptor acquisition workflow was directly adapted from Lempinen, A.'s implementation[77] utilizing molecular features for machine learning

(MORFEUS).[85] MORFEUS allows for easy computing of various steric and electronic descriptors from the CREST generated conformer ensembles as well as featuring an integrated interface for calculating xTB-descriptors and a method for calculating Boltzmann-weighted descriptors. As such, Boltzmann-weighted averages, minimum energy conformers, and minimum/maximum values for all descriptors of all conformers in the ensemble were calculated. Local and global descriptors are calculated directly from CREST outputs with local descriptors being calculated from the $Pd(CO)_2$ fragment, such as the nucleophilicities, electrophilicities, atomic charges, buried volumes, dispersion descriptors, bite angles, and solvent accessible surface areas. In addition, all bond lengths, bond orders, vibrational frequencies, and force constants were calculated for every bond in the fragment. The carbonyl-derived local descriptor values are averaged for the purposes of diminishing errors and emphasizing overarching trends. Calculated global descriptors included HOMO- and LUMO-energies, electron affinities, dipole moments, ionization potentials, universal dispersion descriptors, nucleophilicities, and electrophilicities.

3.1.4 Feature selection and intercorrelation analysis

Although the development of the model has been presented thus far as a comparatively linear chain of events, in reality, several detours were made before arriving at the conclusion to pursue descriptors derived from transition metal complex modelling. Indeed, visualization techniques ranging from bar diagrams to handcrafted databases were applied to the data fairly early on in an attempt to both ease the future feature selection process and guide the development process through the eyes of a chemist.

Simple bar diagram graphs were used to visualize how well the utilized catalyst, base, solvent, substrate, temperature, or reaction time correlated with site-selectivity in the reaction data. For the sake of simplicity, only the conventional and unconventional sites are compared in the final model. Initially, a handful of diarylated reaction outcomes were included in the data and visualized. As there were not enough of these data points to begin with, the decision to exclude them from the data was made, as extrapolating anything meaningful out of a couple data points would have been unlikely. In addition, yields or conversions were not the primary interest of this thesis, thus a simple binary comparison

was sufficient. Including diarylated data into the workflow could be a future point of improvement. While the data didn't yield significant insights as is, it notably emphasized the importance of ligand, base, and solvent. The initial version of the visualization tool only allowed for univariate comparisons. However, a later version allowed the user to specify any number of parameters for multivariate comparisons. The visualization of the reaction data through a bar diagram is presented in **Figure 19**.

Figure 19. A simple bar diagram highlighting the correlation between the utilized solvents and site-selectivity for 3,5-dichloropyridazine. This can be used to quickly discern how homo- or heterogenic the data in literature is.

In addition to the bar diagram visualization tool the total number of different features in both the data sets were analysed manually. The number of different features present in the final data sets of both 2,4-dichloropyiridne and 3,5-dichloropyridazine are presented in **Table 2**.

Table 2. The total number of different features in the data sets of 2,4-dichloropyiridne and 3,5-dichloropyridazine. Note that although 2,4-dichloropyridine has roughly four times the number of reactions, it only utilizes twice as many different ligands in comparison to 3,5-dichloropyridazine.

feature	2,4-dichloropyridine (219 reactions)	3,5-dichloropyridazine (55 reactions)
ligands	8	4
catalysts	21	9
unique bases	17	7
base combinations	20	7
unique solvents	13	7
solvent combinations	23	7
temperatures	19	7
reaction times	37	16
reagents	187	36

The significance of feature selection, the process of choosing a subset of relevant variables, became particularly evident following ligand parametrization. With well over 200 descriptors at hand, eliminating irrelevant and less important features became a necessity for the sake of reducing the model's overall complexity. An appropriate set of features is capable of accelerating the training process, minimizing overfitting as well as increasing the model's accuracy.[59] As stated earlier, it is highly advisable to avoid the use of descriptors that exhibit significant intercorrelations, especially if the primary goal of the modelling is the acquisition of mechanistic insights. Collinear parameters amplify the random noise in the data without the addition of any new information to the model.[8]

Pairwise correlations for the data sets of both 2,4-dichloropyiridne and 3,5-dichloropyridazine were performed to find strongly correlating descriptors. This was visualized through a correlation matrix. Correlations between two different descriptors without the information of the end products site-selectivity were calculated and plotted. The values within each cell of the matrices represent the correlation coefficients between to two intersecting descriptors. A higher value means a higher correlation. Positive

correlations are indicated by a red gradient, while negative correlations are indicated by a blue gradient. A negative correlation implies an inverse correlation between descriptors, that is, an increase in one results in the decrease of another. The information provided by these matrices was used to help with the feature selection process. The individual correlation matrices of 2,4-dichloropyridine and 3,5-dichloropyridazine are presented in **Figure 20** and **Figure 21**, respectively.

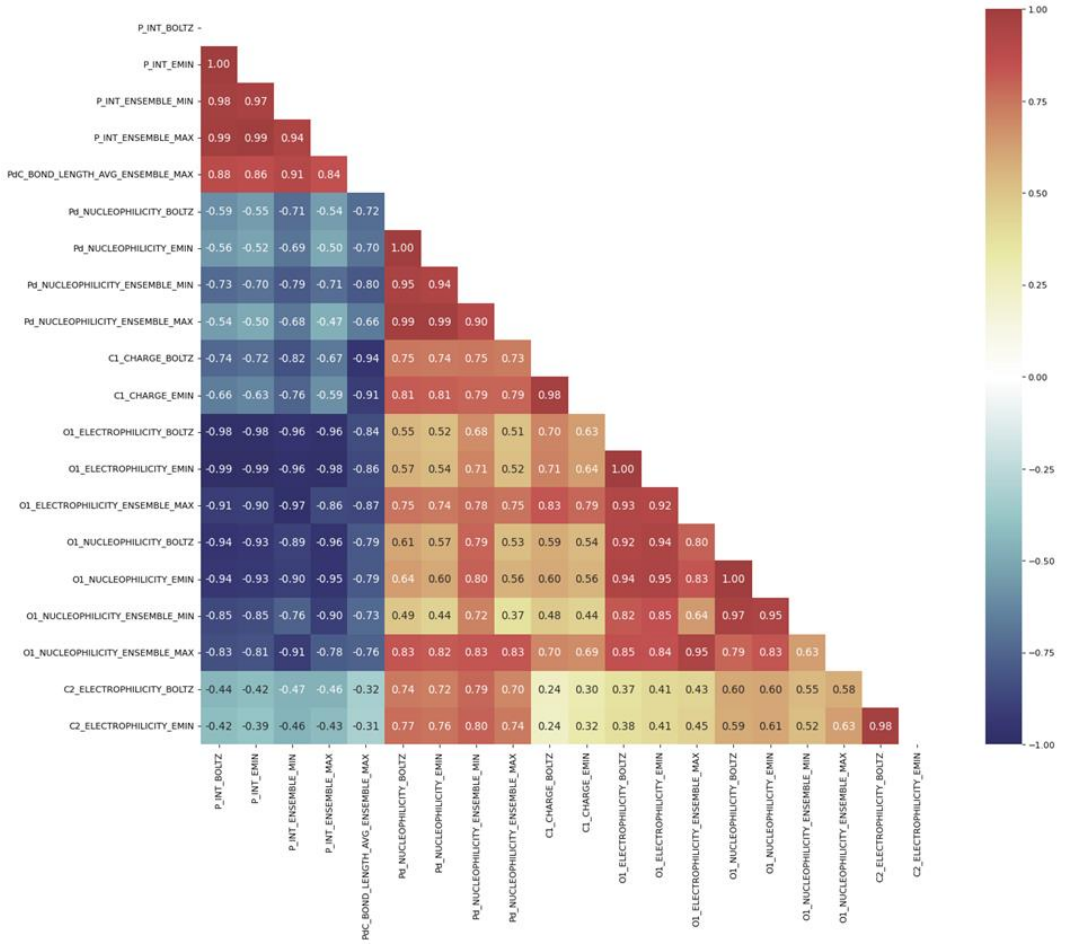


Figure 20. The pairwise correlation matrix for the descriptors of 2,4-dichloropyridine.

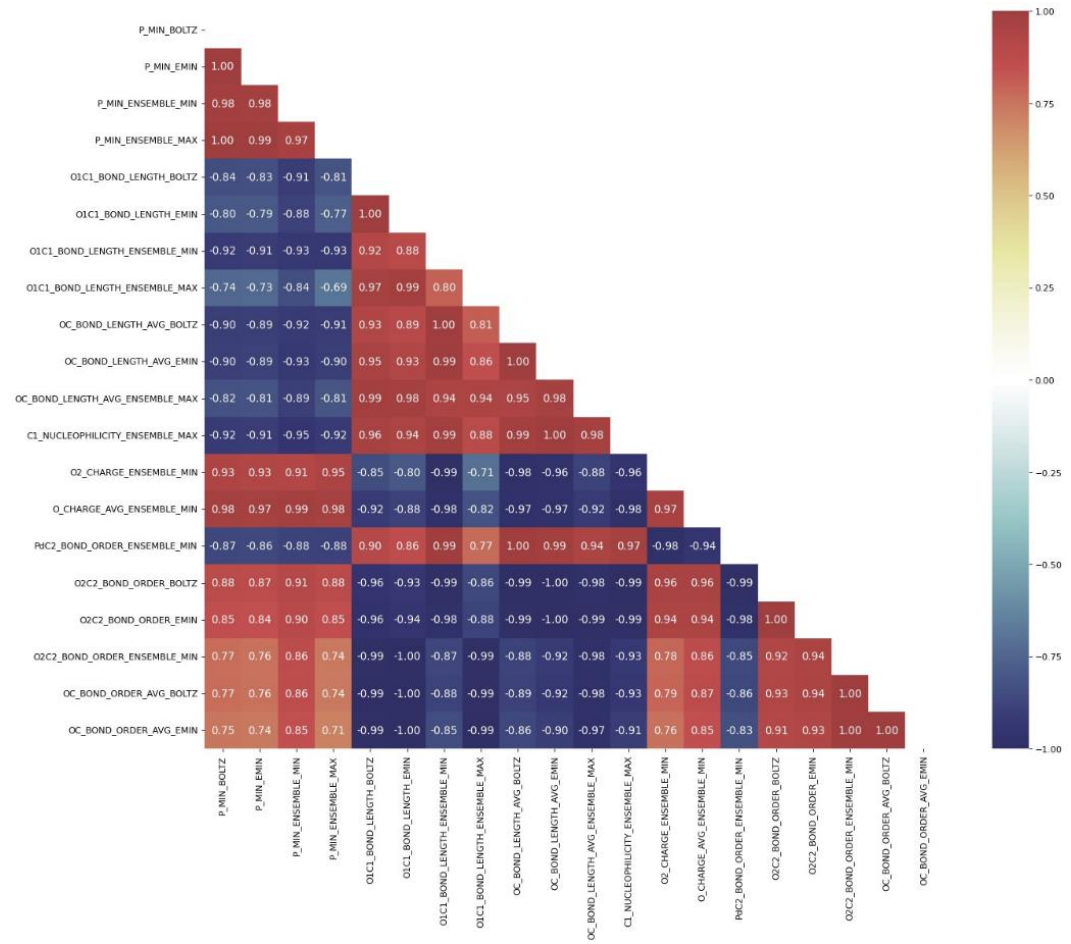


Figure 21. The pairwise correlation matrix for the descriptors of 3,5-dichloropyridazine.

For 2,4-dichloropyridine, the most notable correlating descriptors were for the large part electronic features. While this is nothing out of the ordinary, it nevertheless provides an interesting contrast to the observations seen in literature, where steric effects have been thought to play a larger role in affecting the regiochemical outcomes of SMCC's.[1] Conversely, a greater number of steric features are observed with the correlation matrix of 3,5-dichloropyridazine's data. Nevertheless, the visualization of intercorrelations among the descriptors proved to be a valuable tool in choosing the right features for the initial univariate comparisons performed in the next chapter.

Besides the employment of pairwise correlations for feature selection, feature importances were investigated with a RF classifier using Scikit-learn, an open-source machine learning library for Python.[86] Although RFs are commonly utilized for classification or regression purposes, the tree-based estimators may be also used to compute impurity-based function importance. In the context of DTs, impurity refers to

the likelihood of a data point being incorrectly labelled if it were labelled randomly based on the distribution of labels in a particular node. The mean and standard deviation of accumulation of the impurity decrease within each tree is calculated, which can then be used to discard irrelevant features.[59, 86] Feature importances were calculated for both 2,4-dichloropyridine and 3,5-dichloropyridazine with the addition of base and solvent "descriptors" in the data set. The top twenty features of both substrates sorted by feature importances are presented in **Table 3** and **Table 4**, respectively. Note that steric features are emphasized using bold text.

Table 3. Top twenty 2,4-dichloropyridine descriptors sorted by feature importance.

descriptor	importance (%)
O2_NUCLEOPHILICITY_ENSEMBLE_MIN	7.13
O_NUCLEOPHILICITY_AVG_ENSEMBLE_MAX	5.54
O2_NUCLEOPHILICITY_EMIN	5.44
BASE	4.43
NUCLEOPHILICITY_ENSEMBLE_MIN	3.83
C2_ELECTROPHILICITY_ENSEMBLE_MAX	3.77
C2_ELECTROPHILICITY_BOLTZ	3.74
C2_ELECTROPHILICITY_EMIN	3.73
SOLVENT	3.28
O2C2_FREQ_BOLTZ	2.88
O_NUCLEOPHILICITY_AVG_EMIN	2.88
Pd_NUCLEOPHILICITY_EMIN	2.80
O2_NUCLEOPHILICITY_ENSEMBLE_MAX	2.79
IP_BOLTZ	2.74
IP_EMIN	2.73
O1_NUCLEOPHILICITY_ENSEMBLE_MAX	2.72
NUCLEOPHILICITY_BOLTZ	2.70
O_NUCLEOPHILICITY_AVG_BOLTZ	2.68
Pd_NUCLEOPHILICITY_BOLTZ	1.91

Table 4. Top twenty 3,5-dichloropyridazine descriptors sorted by feature importance.

descriptor	importance (%)
PdC2_BOND_ORDER_ENSEMBLE_MIN	6.00
O2C2_BOND_ORDER_ENSEMBLE_MIN	5.00
O1C1_BOND_LENGTH_ENSEMBLE_MIN	**4.00**
O1_ELECTROPHILICITY_BOLTZ	3.00
O2_CHARGE_ENSEMBLE_MIN	3.00
O1C1_FORCE_CONSTANT_ENSEMBLE_MIN	3.00
O2C2_FORCE_CONSTANT_EMIN	3.00
P_MIN_ENSEMBLE_MIN	**2.00**
P_MIN_ENSEMBLE_MAX	**2.00**
O2C2_BOND_LENGTH_EMIN	**2.00**
OC_BOND_LENGTH_AVG_EMIN	**2.00**
C_VBUR_AVG_ENSEMBLE_MAX	**2.00**
C1_NUCLEOPHILICITY_ENSEMBLE_MAX	2.00
O_ELECTROPHILICITY_AVG_ENSEMBLE_MIN	2.00
PdC1O1_BITE_ANGLE_BOLTZ	**2.00**
PdC1O1_BITE_ANGLE_EMIN	**2.00**
O1C1_FORCE_CONSTANT_EMIN	2.00
OC_FORCE_CONSTANT_AVG_EMIN	2.00
PdC_FREQ_AVG_ENSEMBLE_MAX	2.00
O1C1_FREQ_EMIN	2.00

The RF feature importance analysis appears to highlight descriptors that overlap with those identified in the pairwise correlation matrices. For both substrates, electronic descriptors make up the majority of the important features. However, for the data of 3,5-dichloropyridazine a greater emphasis is placed on the steric features compared to that of 2,4-dichloropyridine. The significance of the base and solvent are also clearly highlighted in the data of 2,4-dichloropyridine, although nothing chemically significant could be concluded from them, due to them being merely statistical representations.

3.1.5 Classification modelling

With the descriptors obtained and most prominent features identified, it was time for the development of the initial models. Various classification models were examined and initially the best descriptors from both of the aforementioned analysis methods were cross-plotted. The explored models included DTs, such as RFs and the AdaBoost algorithm, SVMs, NNs and Gaussian process classifiers (GPCs). These were all implemented using the Python package Scikit-learn.[86] Both data sets were split into training and test sets using a random 80:20 train-test split. The preliminary results from the initial cross-plots for the data of 2,4-dichloropyridazine is presented below in **Figure 22**.

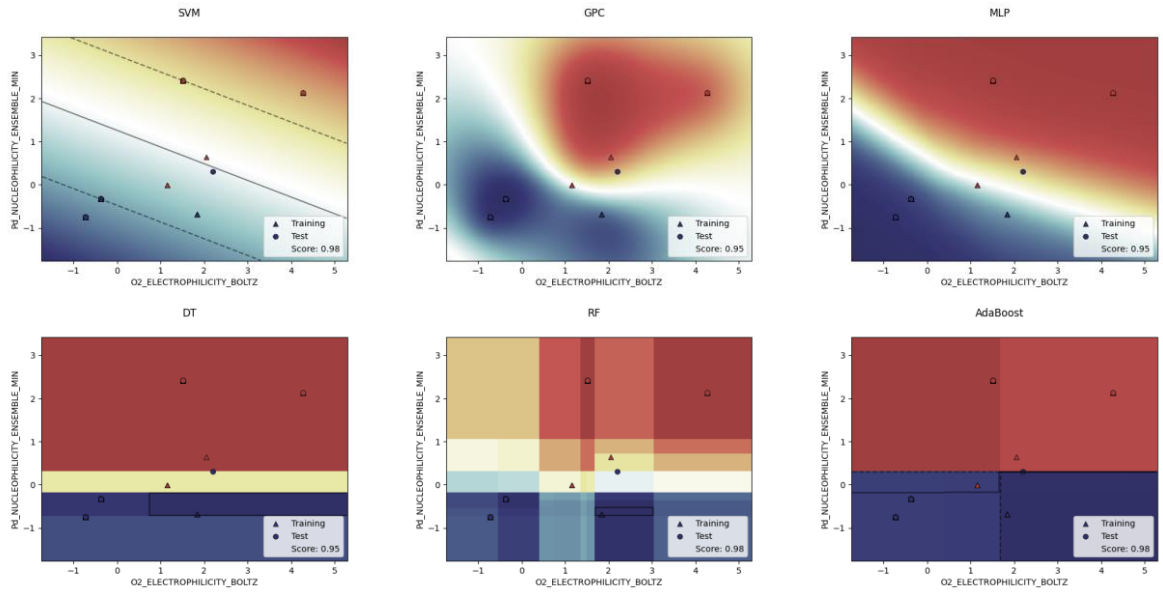


Figure 22. Various classification models producing similar, yet divergent decision boundaries for the data of 2,4-dichloropyridine.

The decision to not yet include any predicted data was made, as the main goal was to first identify the descriptors and the classification model capable of separating the data points from one another in the most obvious manner. In addition, the problem of collinearity and high dimensionality of the data set would have to be first examined. While collinearity might be undesirable for models seeking mechanistic insights, they are far less significant of a problem in the context of classification-oriented models. The concern of collinearity can be reduced even further by performing PCA on the data set. As such, both two- and three-dimensional PCA reduced data sets were initially produced and

evaluated. With the visualization of the decision boundaries for the two-dimensional data being far easier to implement and comprehend, the decision to omit the three-dimensional data altogether was made. Incorporating three- or even four-dimensional data sets into the models could provide better separation of the data points or model reactivity in a more appropriate manner.[87] However, due to time constraints, larger than two-dimensional data sets were not explored further. The examined models utilizing the PCA reduced data set of 3,5-dichloropyridazine is presented below in **Figure 23**.

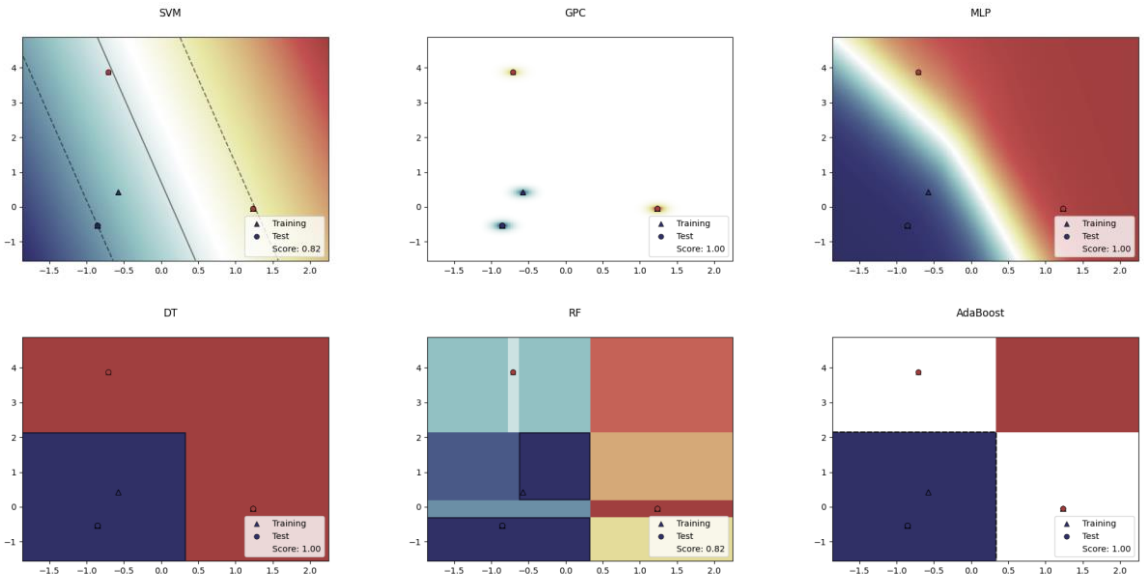
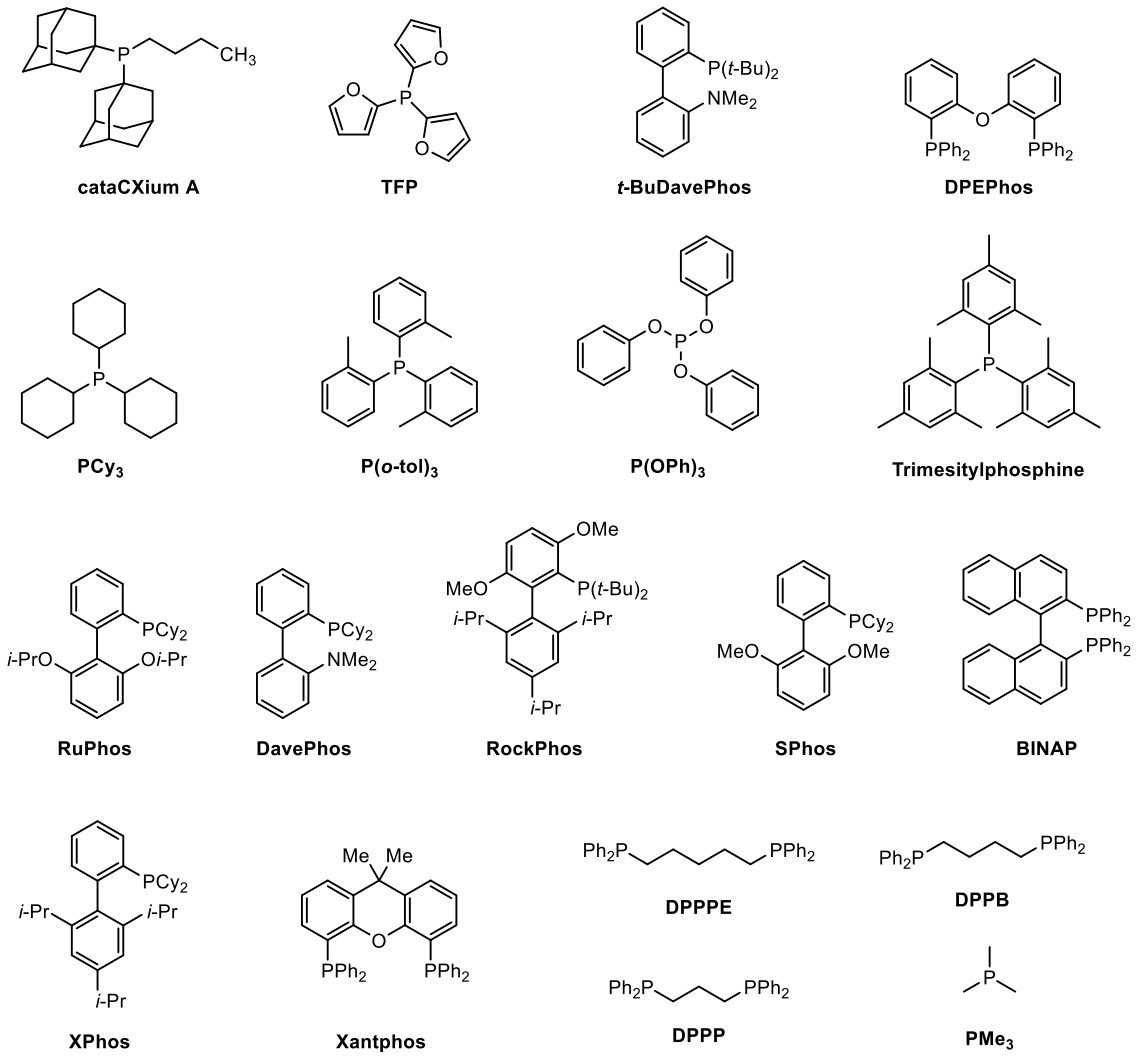


Figure 23. Models utilizing the PCA reduced data sets of 3,5-dichloropyridazine.

Regarding the separability of data points and clarity of decision boundaries, the PCA performed on the data set of 3,5-dichloropyridazine far surpassed that of 2,4-dichloropyridine. However, the likely cause for this was the significantly smaller data set containing overall fewer ligands, and thus, less variance. Acceptable decision boundaries for 2,4-dichloropyridine could be generated by applying the PCA only to the top ten most significant descriptors identified by the RF classifier.

When developing these models, there is always a significant bias towards the attainment of a model that can effectively generalize to the data at hand. However, this does not always guarantee generalizability to unseen data. Additionally, the visualization and the interpretation of data obtained from ML methods is not a straightforward process and represents an ongoing challenge in computer research.[39] While some of the examined

models might not have always produced the most visually appealing results, they could be more in line with reality. The only feasible way to verify this is with more experimentation and real-world data. Regardless, the MLP's ability to separate the data in the most appropriate manner made it the most promising candidate. As such, various commercially available ligands (**Scheme 6**) were modelled in accordance with the previously described workflows and virtually screened with the MLP classification model. It should be noted that at this point the decision to exclusively proceed with the examination of 3,5-dichloropyridazine was made, largely due to time constraints. The resulting virtual screening with MLP is presented in **Figure 24**.



Scheme 6. Virtually screened ligands.

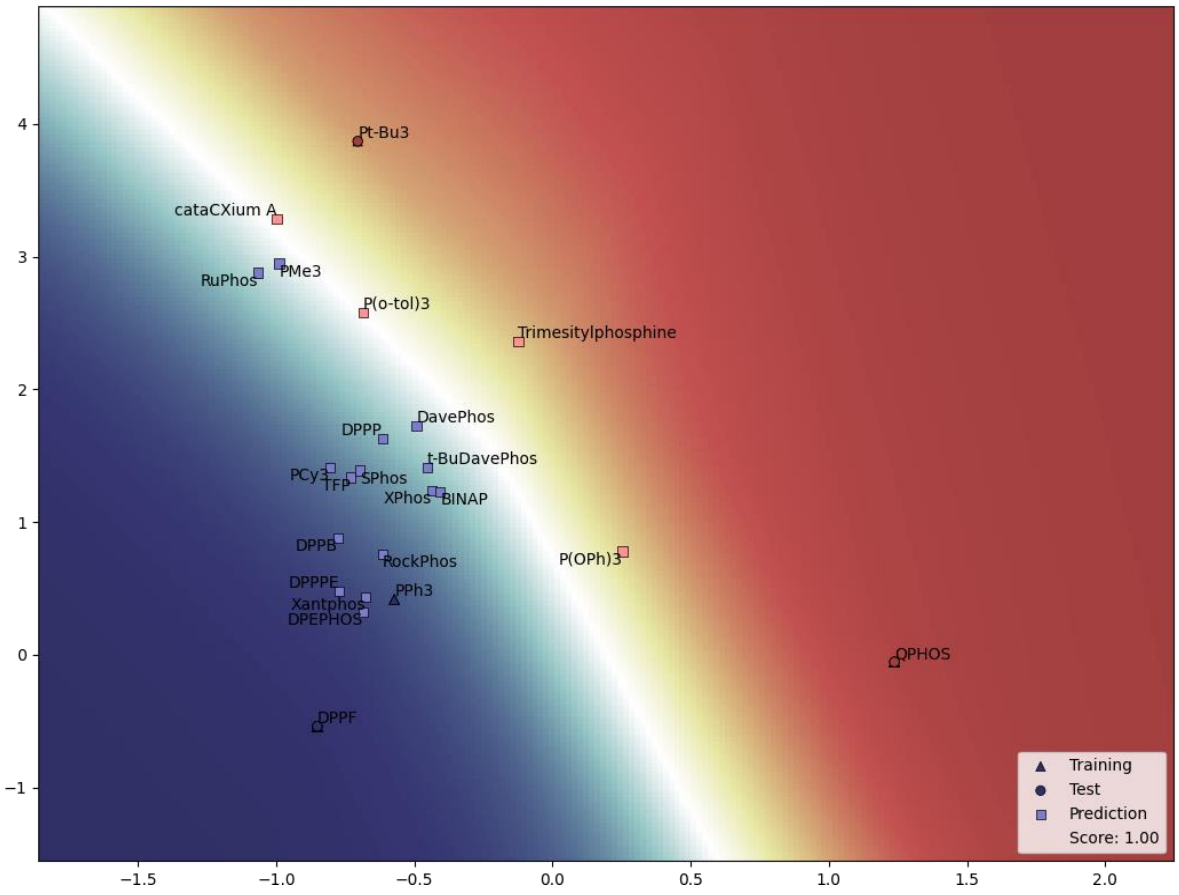


Figure 24. Virtual screening performed using the MLP classifier trained on the data of 3,5-dichloropyridazine. Triangular data points represent the training data, circular points the test data, and square points the predicted data. Ligands highlighted in blue are C3 selective, while C5 selective ligands are highlighted in red.

The outcome of the virtual screening suggests that an overwhelming majority of the ligands are predicted to exhibit C3 selectivity. Only a handful of the predicted C5 selective ligands appear to be in the uncertainty region, possibly hinting at diarylated products. Interestingly, the orthogonal parameters of the PCA seem to group certain ligands based on their physical properties. For instance, many of the Buchwald-type phosphines are clustered together, along with the vast majority of the bidentate ligands.

However, this virtual screening alone was not thought to be enough. Further comparisons with different models were deemed necessary to aid in the selection of the ligands for laboratory testing. As such, a brute force approach to virtual screening was developed by creating models that performed cross-plots with all the descriptors. In essence, this

meant over 20 000 different results per one classification model. These results were then averaged and compared to one another to evaluate their legitimacy and average out any possible errors. Multiple models predicting similar outcomes were thought to be more reflective of reality. The results from the brute force approach to virtual screening are presented below in **Table 5**.

Table 5. Virtual screening results of the different models for all the screened ligands. Each entry represents the ratio of the C3 to C5 outcomes of each model. For instance, for DPEPHOS, 93% of the DT models predicted a C3 outcome.

ligand	DT	RF	SVM	MLP	total
t-BuDavePhos	47 : 53	50 : 50	64 : 36	60 : 40	55 : 45
TFP	57 : 43	59 : 41	71 : 29	64 : 36	63 : 37
cataCXium A	34 : 66	34 : 66	59 : 41	39 : 61	42 : 58
DPEPHOS	93 : 7	94 : 6	81 : 19	97 : 3	91 : 9
PMe_3	55 : 45	57 : 43	67 : 33	51 : 49	58 : 42
RockPhos	61 : 39	59 : 41	73 : 27	69 : 31	66 : 34
$P(o\text{-tol})_3$	27 : 73	31 : 69	59 : 41	45 : 55	40 : 60
SPhos	60 : 40	63 : 37	73 : 27	71 : 29	67 : 33
DPPPE	77 : 23	81 : 19	82 : 18	89 : 11	82 : 18
Xantphos	90 : 10	88 : 12	84 : 16	88 : 12	87 : 13
$P(OPh)_3$	48 : 52	47 : 53	49 : 51	47 : 53	48 : 52
PCy_3	60 : 40	59 : 41	68 : 32	69 : 31	64 : 36
DPPP	58 : 42	61 : 39	73 : 27	69 : 31	65 : 35
DPPB	59 : 41	63 : 37	74 : 26	74 : 26	68 : 32
Trimesitylphosphine	17 : 83	21 : 79	53 : 47	35 : 65	32 : 68
RuPhos	45 : 55	46 : 54	73 : 27	54 : 46	54 : 46
XPhos	61 : 39	60 : 40	65 : 35	65 : 35	63 : 37
DavePhos	44 : 56	48 : 52	63 : 37	58 : 42	53 : 47
BINAP	63 : 37	64 : 36	70 : 30	71 : 29	67 : 33

3.2 Laboratory testing and re-training

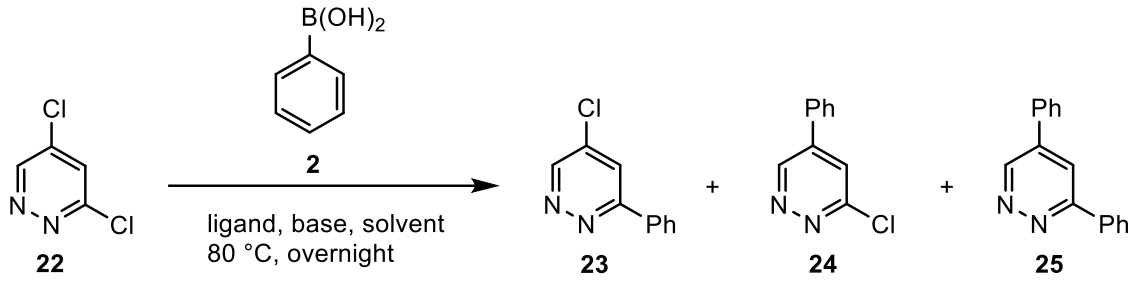
3.2.1 Synthesis procedures

In addition to the ligands already present in the data set, four more were to be tested in the laboratory. Several factors would affect the selection process of these ligands. Namely, the results of the virtual screening, structural diversity, and previous occurrences in literature. Ideally, some of the ligands would have previously documented results, while others would be novel and untested in this context. Out of the virtually screened ligands, cataCXium A, TFP, DPEPhos and *t*-BuDavePhos fit the aforementioned criteria best and were chosen for further examination. Ligands included in the initial data set, that is PPh_3, P*t*-Bu_3, QPhos, and DPPF, were slated for testing with the purpose of model validation.

Synthesis procedures would be directly based on the work conducted by Dai and co-workers.[4] A combined total of 32 reactions, each involving four reactions per ligand, were conducted. This meant a total of two reactions per different base and solvent combination. In addition, two test reactions were conducted to ascertain whether Dai and colleagues had performed their optimized reactions using either solvent or a solvent mixture, as this was rather vaguely stated in their work. As a result of the conducted test reactions, it became evident that Dai and co-workers had used a solvent-water mixture. Thus, all subsequent reactions were conducted using similar solvent systems. Reagents and solvents were utilized without additional purification or drying processes. However, stock solutions of the reagents were prepared with a solvent-to-water ratio of 4:1. The test reactions were carried out using a microwave, whereas the rest of the reactions were stirred and heated using a hot plate. Structural verification of the crude products involved liquid chromatography-mass spectrometry (LC-MS) analysis, where the product mass peaks were integrated and compared against the integrated peaks of the side products. NMR would later be employed to confirm the structures of the purified products.

The general synthesis was as follows: A vial equipped with a stir bar was placed under a gentle nitrogen flow and charged with 3,5-dichloropyridazine (**22**, 0.200 mL, 0.2 mmol, 1 M sol in 1,4-dioxane, 1 eq.), phenylboronic acid (**2**, 0.200 mL, 0.2 mmol, 1 M sol in 1,4-dioxane, 1 eq.), the base (0.200 mL, 0.6 mmol, 3 M sol in H_2O, 3 eq.), and 0.320 mL of the degassed solvent. After a brief period, Pd(II)$(OAc)_2$ (0.08 mL, 0.002 mmol, 0.25 M sol in

1,4-dioxane, 0.1 eq.) was added along with either 0.1 or 0.2 equivalents of the ligand (ligand was added as a solid), depending on its denticity. A 1:1 ligand/Pd ratio was used for bidentate ligands and a 2:1 ligand/Pd ratio for monodentate ligands. The vial was then purged with nitrogen for an additional minute before closing. The resulting reaction mixture was stirred and heated at 80 °C for 16 hours using an aluminium heating block. Analysis was performed right after the reaction with LC-MS directly from the reaction mixture. The results of the performed reactions are presented below in **Scheme 7**.



entry	ligand	base	solvent	23	24	25
1	DPPF	Cs_2CO_3	dioxane/H_2O (4:1)	59	5	36
2	cataCXium A	Cs_2CO_3	dioxane/H_2O (4:1)	19	47	34
3	PPh_3	Cs_2CO_3	dioxane/H_2O (4:1)	71	6	23
4	P*t*-Bu_3	Cs_2CO_3	dioxane/H_2O (4:1)	9	56	35
5	DPEPhos	Cs_2CO_3	dioxane/H_2O (4:1)	42	13	45
6	*t*-BuDavePhos	Cs_2CO_3	dioxane/H_2O (4:1)	26	35	39
7	TFP	Cs_2CO_3	dioxane/H_2O (4:1)	42	8	50
8	QPhos	Cs_2CO_3	dioxane/H_2O (4:1)	0	37	63
9	DPPF	KF	toluene/H_2O (4:1)	51	29	20
10	QPhos	KF	toluene/H_2O (4:1)	6	83	11
11	cataCXium A	KF	toluene/H_2O (4:1)	16	63	21
12	PPh_3	KF	toluene/H_2O (4:1)	46	42	12
13	DPEPhos	KF	toluene/H_2O (4:1)	28	50	22
14	TFP	KF	toluene/H_2O (4:1)	13	66	21
15	*t*-BuDavePhos	KF	toluene/H_2O (4:1)	9	49	42
16	P*t*-Bu_3	KF	toluene/H_2O (4:1)	12	52	36

Scheme 7. Reaction conditions and results of the performed reactions. Entries are an average of two performed reactions.

Based on the performed reactions, it is evident that choosing the correct combination of base and solvent is as equally important as selecting the appropriate ligand. The general trend appears to be that in the presence of Cs_2CO_3 and dioxane/H_2O (4:1), significantly larger quantities of selective coupling at the C3 position and the formation of the

diarylated product is observed. Conversely, with KF and toluene/H_2O (4:1) used as the base-solvent combination, there is an overall decrease in diarylation, and a greater degree of selective coupling is observed at the unconventional C5 position. This even appeared to be the case when using traditionally C3 selective ligands, such as PPh_3 or DPPF. The most interesting example was TFP, where an almost complete inverse in selectivity is observed with the KF and toluene/H_2O (4:1). As such, the line between conventional and unconventional selectivity blurs.

3.2.2 Purification and characterization

As stated earlier, the relative conversions of the product isomers for each performed reaction were determined with LC-MS analysis immediately following the reactions. In addition, the presence of additional by-products and impurities, such as biphenyl, were verified with the aid of LC-MS. The source of biphenyl is likely from two phenylboronic acids repeated participation in the transmetalation step, which is a common side reaction in Pd mediated couplings.[88] An example spectra for a reaction involving P*t*-Bu_3 in Cs_2CO_3 and dioxane/H_2O (4:1) is presented in **Figure 25**. The remaining LC-MS spectra are included in the supporting materials.

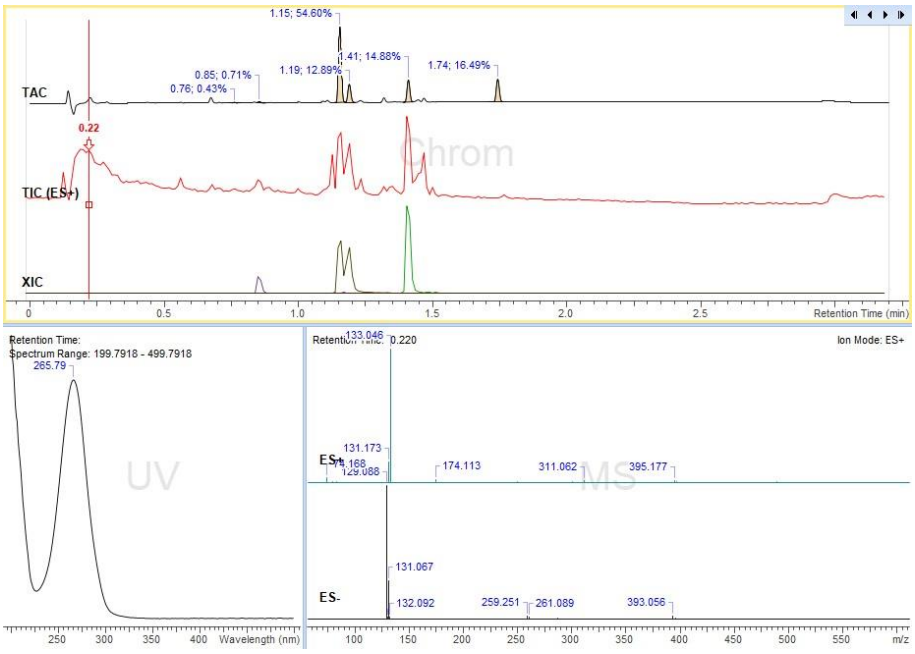


Figure 25. LC-MS spectra for P*t*-Bu_3 in Cs_2CO_3 and dioxane/H_2O (4:1) demonstrating the ease of identifying different product isomers and by-products.

While it was assumed that the compounds with the retention times of around 1.14 to 1.20 minutes were the two main product isomers **24** and **23** respectively, their structures still had to be verified using NMR spectroscopy. Based on the mass spectra, the compound with the retention time of approximately 1.40 had to be the diarylated product **25**. As diarylation was not of particular interest nor the main focus of this thesis, only the two regioisomers were analysed. The reactions with the best C3 and C5 conversions were combined separately for further purification and structure verification. The product isomers were then separated utilizing a standard high-performance liquid chromatography (HPLC) technique. The combined retention time fractions of both C3 and C5 isomers were concentrated in vacuo and dried.

The products were then characterized by ^{1}H-NMR, ^{13}C-NMR and nuclear Overhauser effect spectroscopy (NOESY). In addition, correlated spectroscopy (COSY), heteronuclear single quantum coherence (HSQC) and heteronuclear multiple bond correlation (HMBC) were performed, but their spectra were not deemed necessary for the purposes of structural elucidation. Due to differing proton-proton interactions arising from structural differences, an extra peak is detected in the NOESY spectra for **24** when compared to **23**. This made structural ascertainment a relatively straightforward process. The compounds with retention times of ~1.15 and ~1.20 were confirmed to be **24** and **23**, respectively. All NMR spectra and chemical shifts are reported in the supporting materials section.

3.2.3 Re-training

After having performed the first half of the reactions, the model was re-trained for the first time. Intercorrelation and RF classification analysis yielded largely comparable results to those obtained from the pure literature data. The resulting MLP classifier trained on this new data also produced similar results. However, after performing the latter half of the reactions and the subsequent second re-training, the relevance of the base and solvent became distinctly emphasized in the analysis of relevant features. Again, it should be stated that the two “descriptors” are merely statistical representations.

Initially, decision boundaries comparable to those of 2,4-dichloropyridine were obtained when applying all relevant descriptors with PCA. Similarly, reducing the number of descriptors to the top twenty or so resulted in more visually appealing boundaries. This time around the predictions were more heavily skewed toward the unconventional product. The extent to which this model reflects reality remains uncertain for the time being. Virtual screening performed with the finalized MLP classification model is presented in **Figure 26**.

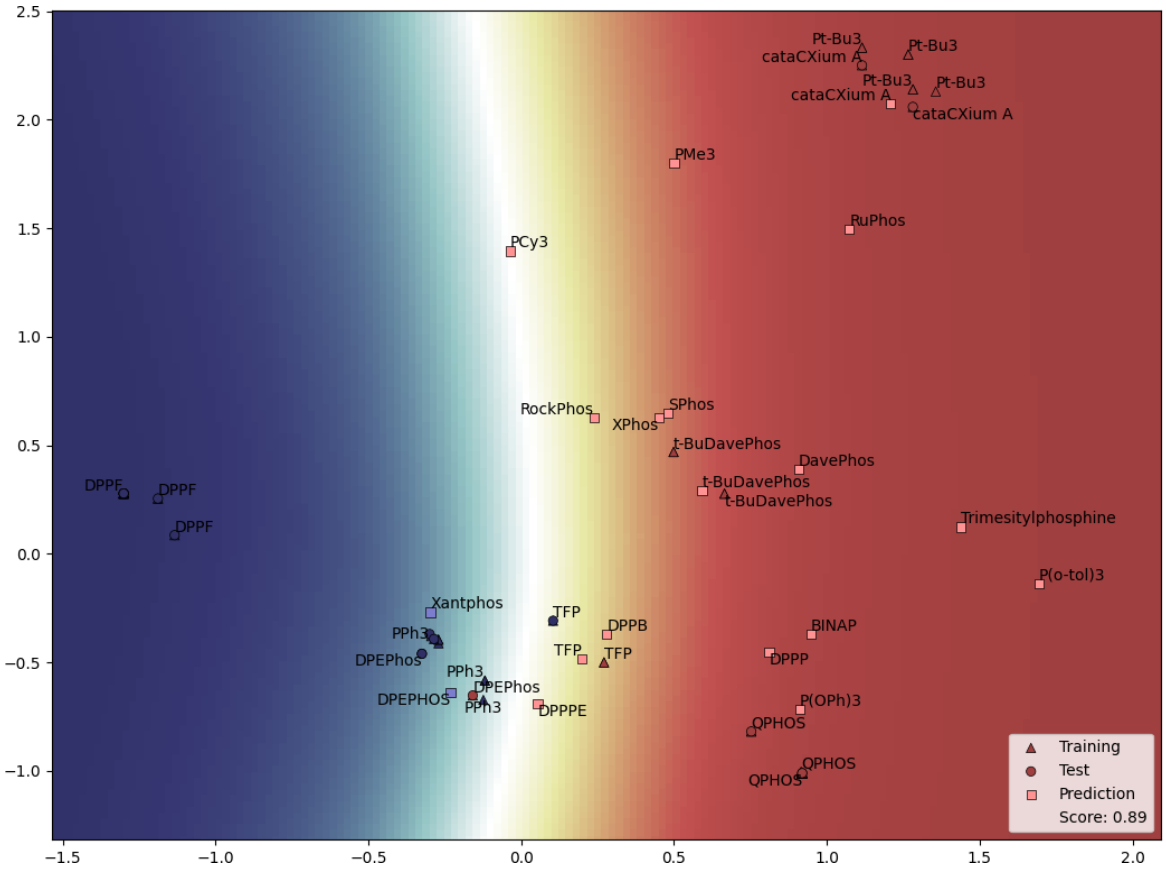


Figure 26. Virtual screening performed with the final MLP classifier. PCA's orthogonal parameters are a result of the top twenty most relevant features according to the RF classification analysis. Note the significantly larger number of ligands classified as C5 selective this time.

4 Discussion

This chapter serves both as an introspect of the work performed and a discussion on the results and future outlook of the various aspects discussed earlier in this thesis. As a whole, this was a tremendous learning process for me. Although I had a general understanding of catalytic chemistry and an extensive background on programming, the sheer complexity of the seemingly simple task of predicting the regiochemical outcomes of *N*-heteroarenes came to me as a surprise. Nevertheless, with the right amount of guidance form Orion corporation and thorough research into both computational and catalytic chemistry, I was able develop the predictive models I was tasked with, along with the workflows required for developing them. The groundwork laid here should allow for faster and easier development of comparable models in the future.

That being said, there were many aspects of this work that went as planned as well as those that could have gone significantly better. While I have already briefly reflected on some of these issues, I have thought it prudent to perform a more comprehensive analysis to prevent repeating similar mistakes down the line. First, the data-acquisition. With no prior experience in the utilization of reaction databases, much of the data mining process could have been executed more effectively. Utilizing Reaxys instead of SciFinder would have indeed been easier, as Reaxys allows the user to download the reaction data directly to a spreadsheet. While SciFinder offers somewhat similar functionality, it would have been considerably more challenging to implement.

Furthermore, contrary to my assumption, much of the reaction data is not actually present within the database itself but hidden away in the publications supplementary information. As to why these have not been catalogued, I can only assume the databases being constructed primarily with simplicity and the chemist in mind. However, the inclusion of negative examples could also benefit chemists by allowing them to discern between better and worse reaction conditions more quickly. The current unfortunate reality is that only the best and most optimized reactions get reported, leaving out much of the essential data required for the construction of accurate models. As much as I detest performing unnecessary manual labour, further investigation on the reactions should have been ultimately conducted.

On a similar note, the problem of diarylation should have also been investigated from the start. While some the diarylated reaction outcomes were initially included in the data set, the rationale for omitting these was their scarcity in the database. However, as is evident from the lack of investigation performed on the publication's supplementary information, much of the valuable information was ultimately overlooked. The incorporation of yields and conversions of the different reaction products is the next obvious step in improving the performance of the model.

Regarding the palladium complex modelling, comparatively little thought was put into the validity of utilizing $Pd(CO)_2$ fragments. As research has shown, palladium carbonyl compounds are not particularly stable.[81] Furthermore, in comparison to $Ni(CO)_2$, the average C-O bond length in palladium carbonyl compounds is below the threshold where both σ- and π-bonding are most in balance. Given that the contribution from π back-bonding is significantly smaller in these compounds, its impact on the other properties of the complex should likewise be diminished.[76, 82, 83] As such, the reasoning behind the work of Lempinen, A. is not necessarily applicable here.[77] The extent to which the steric effects of additional ligands coordinated to the palladium affect the carbonyl ligands remains unknown. A more rigorous analysis on the credibility of the modelled complexes should have been carried out. My lack of prior knowledge and experience in this field largely influenced the decision to blindly follow this type of modelling workflow. However, with my current level of knowledge, the development of more novel workflows is definitely within the realm of possibility.

Based on the work conducted by Ahneman and co-workers, it should be evident that a mechanistic analysis of the descriptors employed in nonlinear models is not a straightforward process.[66] Performing a similar evaluation of the relative importances reveals the significance of the base-solvent combination in the regiochemical outcome of the reactions. However, this observation was already evident from the results of the laboratory experiments. If the RF classifier had not emphasized any of these features, then the credibility of this approach would have been put into question. A similar ratio of other electronic and steric descriptors is observed when compared to the initial analysis performed on pure literature data. The results of the feature importance analysis are presented in **Table 6**.

Table 6. The final top twenty 3,5-dichloropyridazine descriptors sorted by feature importance.

descriptor	importance (%)
SOLVENT	5.93
O1_ELECTROPHILICITY_ENSEMBLE_MIN	5.90
O2C2_FORCE_CONSTANT_EMIN	5.89
BASE	5.57
P_MIN_ENSEMBLE_MAX	**4.83**
PdC2_BOND_ORDER_ENSEMBLE_MIN	4.80
O1_ELECTROPHILICITY_BOLTZ	4.47
O2C2_BOND_LENGTH_EMIN	**3.42**
O1_ELECTROPHILICITY_EMIN	3.32
O_ELECTROPHILICITY_AVG_EMIN	3.06
OC_BOND_LENGTH_AVG_EMIN	**2.65**
O1C1_BOND_LENGTH_ENSEMBLE_MIN	**2.57**
OC_FORCE_CONSTANT_AVG_EMIN	2.47
O2_CHARGE_ENSEMBLE_MIN	2.39
O1C1_BOND_LENGTH_BOLTZ	**1.81**
PdC1O1_BITE_ANGLE_BOLTZ	**1.80**
PdC1O1_BITE_ANGLE_EMIN	**1.73**
O2C2_BOND_ORDER_EMIN	1.66
O2C2_FREQ_EMIN	1.66
O1C1_FREQ_BOLTZ	1.57

Arguably, the greatest limitation imposed by the representation of bases and solvents by purely statistical means is their inability to offer any concrete mechanistic insights. Instead of multi-hot tables or decimal encoding, molecular descriptors, either experimental or computational, should have been ultimately employed. As previous studies have shown, solvents play a role in stabilizing organometallic catalysts, activating precatalysts, and modulating the reactivity of acids and bases.[26] Thus, solvent descriptors ought to capture the various effects on both the catalyst and the base. Descriptors, such

as the dielectric constant, dipole moment, viscosity, and donor numbers have been utilized in the modelling of solvents before.[89] Incorporating solvent descriptors and solvation modelling[90] to this workflow in the future should be fairly trivial.

If the driving force behind unconventional site-selectivity is indeed the 12-electron PdL and its displacement-type mechanism, then what can be hypothesized from the available data? As solvent properties have been shown to influence the preferred mechanism for oxidative addition, polar solvents seem to favour this displacement-type pathway more.[1] Where the previous rationale was that the solvent polarity stabilizes the transition state structures of oxidative addition involving charged palladium species, the actual role of solvents may be significantly more complex. Reeves and co-workers investigated the matter of a solvent induced switch in selectivity with chloroaryl triflates when using polar solvents. While not a direct one-to-one comparison, the system is similar to the dihaloheteroarenes examined in this thesis in that one site is conventionally more reactive than the other. The authors investigated various polar and nonpolar solvents and their effect on site-selectivity. The divergent results suggest that oxidative addition may not proceed through the same mechanism in all polar solvents, thus making it clear that the dielectric constant alone does not trend with site-selectivity.[26]

From the results of the conducted laboratory experiments, it would be reasonable to assume that for PPh_3 the active species would be a PdL complex in the presence of KF and toluene/H_2O (4:1). Both computational and experimental studies support the notion that a monoligated species could be the active one during catalysis.[91] Conversely, other studies have concluded that the monophosphine species has a lifetime too short to be accessible in solution due to the solvent molecules coordinating to palladium. Even with a nonpolar solvent such as toluene with weak coordinative properties, the palladium rapidly forms a π-complex with at least one of the solvent molecules. For stronger coordinating solvents, the occurrence of solvent coordinated species will be even more likely. However, in the case of more sterically hindered ligands, such as P*t*-Bu_3, the mechanism responsible for the formation of solvent-coordinated species may not be operational.[92] Similar observations were made in the laboratory experiments with P*t*-Bu_3, as its outcomes were hardly affected by the different base-solvent combinations.

Although the role of the base in the catalytic cycle has been traditionally attributed to facilitating the transmetalation step[88], it has been also proposed that the presence of a strong base, such as sodium isopropoxide, contributes to the formation of monoligated species.[91] However, this alone cannot explain the observed divergence in site-selectivity, as Cs_2CO_3 is a stronger base than KF. To what extent do the properties of the base and solvent influence the formation of the unconventional product? Attempting to draw any meaningful conclusions from the low sample size of only two is nigh impossible. Further experimentation with multiple different base-solvent combinations would have to be conducted. Additionally, the effect of aqueous conditions would have to be examined.[93]

While the role of the base and solvent has been emphasized thus far, the importance of the ligand cannot be overlooked. After all, this was a ligand-controlled approach. Examining the remaining descriptors reveals the preference in selectivity imparted by the ligands to be a combined result of both steric and electronic effects. From this table alone, it is hard to discern which features are from the ligands properties and which are from the solvent and base. However, by comparing the table to earlier results where the role of the solvent and base was not as pronounced, one should be able to identify overlapping features. These common descriptors should provide a general impression of the characteristics imparted solely by the ligands. Most prominent features appearing in both tables are bond lengths, bite angles, and electrophilicity of the carbonyl molecules. With steric and electronic descriptors being emphasized once again, this supports the conclusions drawn in pre-existing literature.[6, 22]

The highlighted bite angles and bond lengths should be indicative of the steric effects of other ligands affecting the carbonyls. Similarly, the electronic properties of the ligands should affect those of the carbonyls.[82, 83] Based on the results of the experiments, the steric bulk and electron richness appears to largely contribute to unconventional site-selectivity. Once more, these findings have already been reiterated numerous times in the existing literature.[1] Although QPhos should fit this category perfectly, the base-solvent combination had a significantly greater effect on the degree of diarylation. Perhaps this can be partly attributed to the ferrocene structure of the monodentate ligand. Furthermore, smaller and less electron rich ligand's regiochemical outcomes appeared to be more sensitive to the utilized base and solvent, as divergent results were observed

with both PPh_3 and TFP. In the case of TFP, the smaller ligand size and the 2-furyl substituents acting as more electron withdrawing on the phosphorus nuclei[94] likely contributed to it being more susceptible to the base-solvent combination. For more accurate predictions, the electron richness (pK_a values) of the ligands would have to be calculated and included in the workflow.[91, 95]

Deriving any meaningful conclusions from the utilized bidentates is hard, as only two ligands were tested. Where DPPF was hardly affected by the base and solvent, DPEPhos yielded both the conventional and unconventional products as the primary outcomes under different reaction conditions. For a more comprehensive understanding of the role of ligand properties in site-selectivity, a broader array of ligand types will have to be tested out in the future.

Although additional data is always better for the development of ML models, the previously highlighted issues could also be indicative of the limitations imposed by the currently available chemical descriptors. Needless to say, the task of developing new and improved descriptors is challenging. The myriad of mechanistic and structural diversity encountered in organic reactions, as well as their inherent susceptibility to various reaction conditions, must all be accounted for. Even minor factors, such as daily irreproducibilities, can significantly alter the reaction outcomes. In order to capture all these effects, descriptors that account for the connectivity of molecular graphs, stereoelectronic properties, and the three-dimensional conformations of molecules would have to be developed.[96]

The challenge not only lies in the descriptors themselves but also in the insufficient number of literature examples on which to train the ML models. With an order of ten million known reactions and as many as 20,000 to 30,000 distinct reaction types, the statistics for learning are generally insufficient. This coupled with the lack of comprehensive stoichiometric information in literature and the limitations of our current descriptors, it is no wonder that present ML methods are unable to perform well in predicting reaction outcomes.[96] A potential solution to this problem could entail the development of a larger ML model comprising of smaller, specialized models designed specifically for individual reaction types. However, when accounting for the number of

reactions required for each individual model's development, this quickly becomes impractical, if not impossible. Without major developments in the current descriptors-based models, I anticipate a broader adoption of more generalized approaches, such as LLMs, in the future.

Lastly, I want to explore the potential that LLMs such as ChatGPT have for organic chemistry. Despite some of the recent successes, it however remains that LLMs struggle to incorporate any physical or chemical intuition into their outputs. This is likely due to the limited amount of chemical literature it has been trained on. Many, if not most, scientific articles and results have not been seen by ChatGPT. LLMs are currently limited to echoing the insufficient number of literature results associated with the keywords provided by the user. This lack of training data may also pose a challenge for the model in inferring the relationship between two chemical concepts. As such, ChatGPT currently serves as a tool to make more effective utilization of the knowledge scientists have accumulated over the years.[71, 97]

While the strength of LLMs lies in their ability to be trained and utilized on the basis of questions formulated in natural language, the most significant limiting factor in learning may be the way chemical data is currently presented in literature. LLMs work based on the tokenization of words. That is, the model has learned the correlations between text fragments and uses those correlations to produce coherent sentences in the context given by the user.[97] With purely verbal input, attempting to represent the correlations between two- or even three-dimensional molecular structures of reactions is difficult. As such, my suggestion involves teaching the models about the correlations of the individual molecular fragments between reaction mechanisms. A tokenization of the language of organic chemistry if you will.

To some extent, this has already been realized through the utilization of SMILES and molecular fingerprints.[39, 98] However, the current problem with SMILES is that LLMs have been trained on natural text instead of line representations. Indeed, ChatGPT has been reported to hallucinate the SMILES of the compounds it has been asked.[71, 97] This is an issue that resides either within the representation itself or simply in the lack of training. I believe the most effective approach going forward would be to begin tokenizing the

complete reaction mechanisms with curly arrows included and then explore whether the model is able to infer any correlations within them. The simplest method would involve integrating this approach into an existing LLM. However, I believe that a completely new model designed specifically for organic chemistry would prove more advantageous in the long term. This could provide the model with an inherent, deeper understanding of the underlying reasons of how and why the mechanisms function, as opposed to simply mirroring the results of pre-existing literature.

5 Conclusion

The goal of this thesis was the development of predictive models for ligands influencing the regiochemical outcomes of six-membered *N*-heteroarenes, along with the accompanying tools and workflows. In this regard, the objective was reached, and the initial model proved to accurately predict the outcomes of the four experimentally verified out-of-sample ligands. However, after performing the second round of experimentations, it quickly became apparent that selecting the correct base-solvent combination is just as important as choosing the appropriate ligand.

For 3,5-dichloropyridazine, the general trend appeared to be that in the presence of Cs_2CO_3 and dioxane/H_2O (4:1), selective coupling at the C3 position dominated along with higher levels of diarylation. Conversely, when KF and toluene/H_2O (4:1) employed as the base-solvent combination, a general reduction in diarylation was observed, along with a higher degree of selective coupling at the unconventional C5 position. An interesting switch in selectivity was observed with multiple initially C3 selective ligands, such as PPh_3 and TFP. The mechanistic rationales for these were explored earlier in the discussion.

The extent to which the properties of the base and solvent influence in the formation of the unconventional product remains unknown. While the descriptors emphasize the importance of steric and especially electronic features, meaningful conclusions could not be derived from a sample size of two different base-solvent combinations. Further experimentation would have to be conducted to clearly differentiate between descriptors correlating with the preference in selectivity imparted by the ligands and those associated with the solvent and base. Solvent and base descriptors along with solvation modelling could also improve both the prediction accuracy and the attainable mechanistic insights of the models.

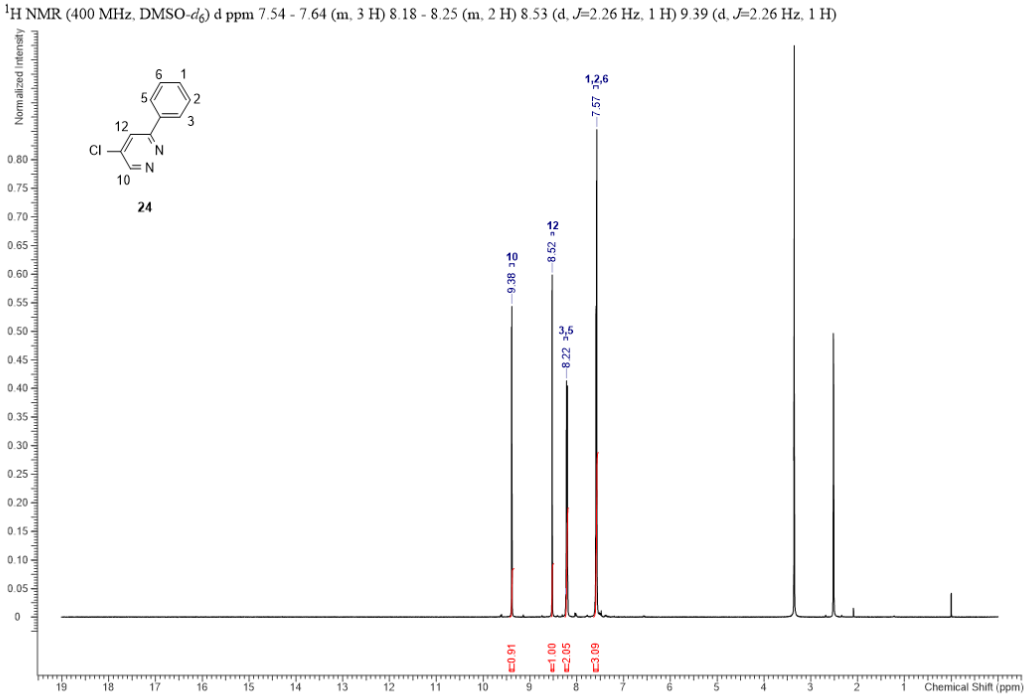
Due to time constraints, the experimental validation of 2,4-dichloropyridine model was not conducted. As such, only a miniscule region of the reaction space of *N*-heteroarenes was explored in the end. For a more comprehensive understanding of the issue of preferential site-selectivity of *N*-heteroarenes, multiple substrate classes along with

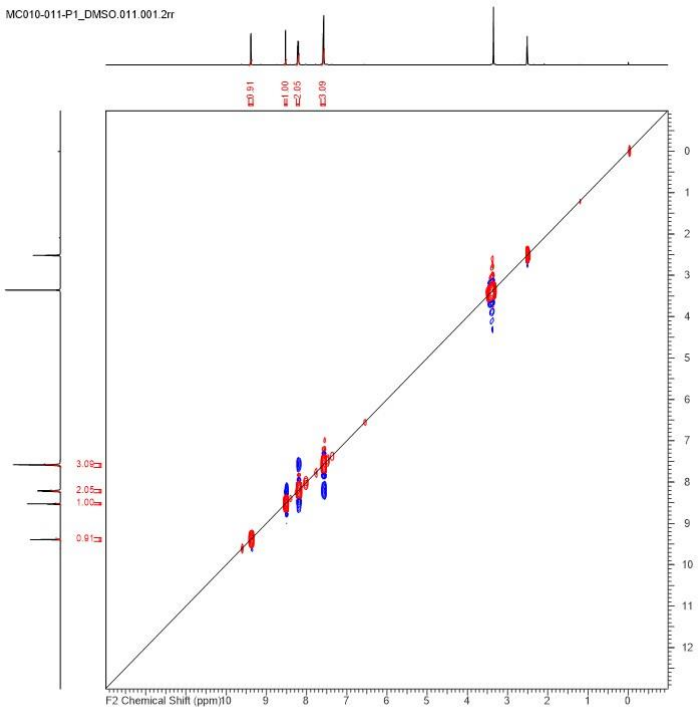
different reaction conditions would have to be investigated. However, careful consideration must be given to the selection of these various parameters, as otherwise, the reaction dimensionality quickly becomes and experimental impossibility.

The groundwork laid with this work should allow for faster and more accessible development of comparable descriptor-based models in the future. The expandability and adaptability of the developed models and workflows allows them to be employed in the modelling of a wide range of cross-couplings, even beyond the palladium catalysed SMCCs that were at center of this thesis. While many aspects of this work could benefit from the improvements discussed earlier, I would consider the overall endeavour a success. This is particularly noteworthy, given that it was my first-time ever conducting research of this nature.

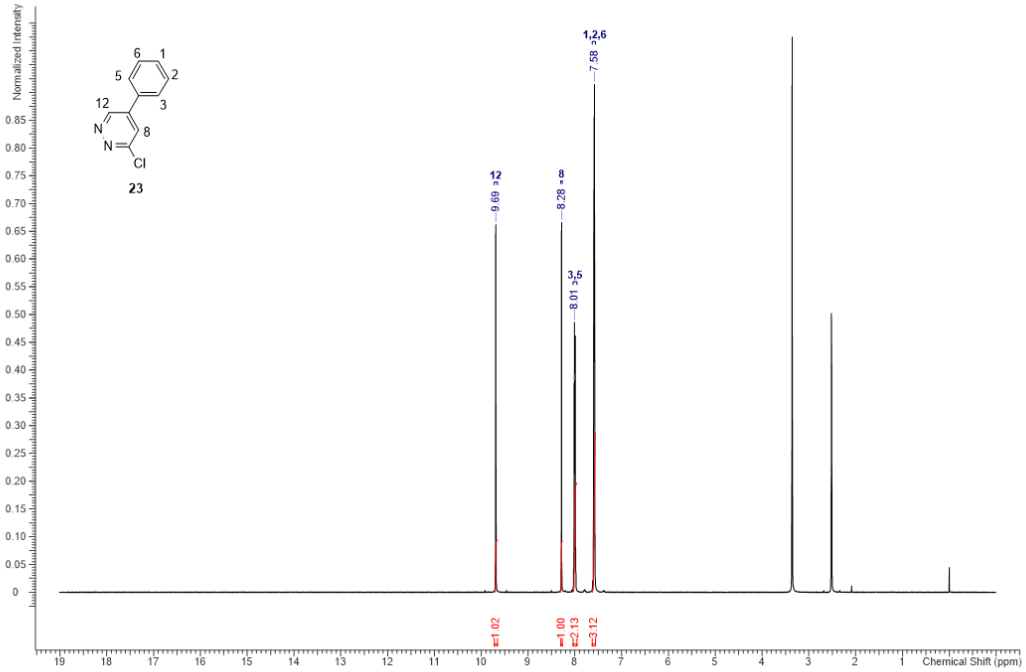
Supporting material

The original data for these spectra was provided by and is held by the Orion Corporation.

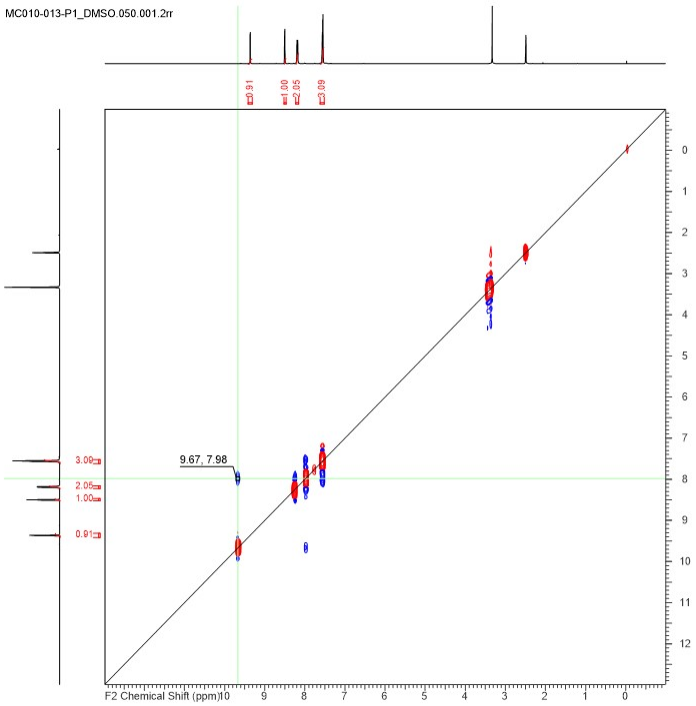


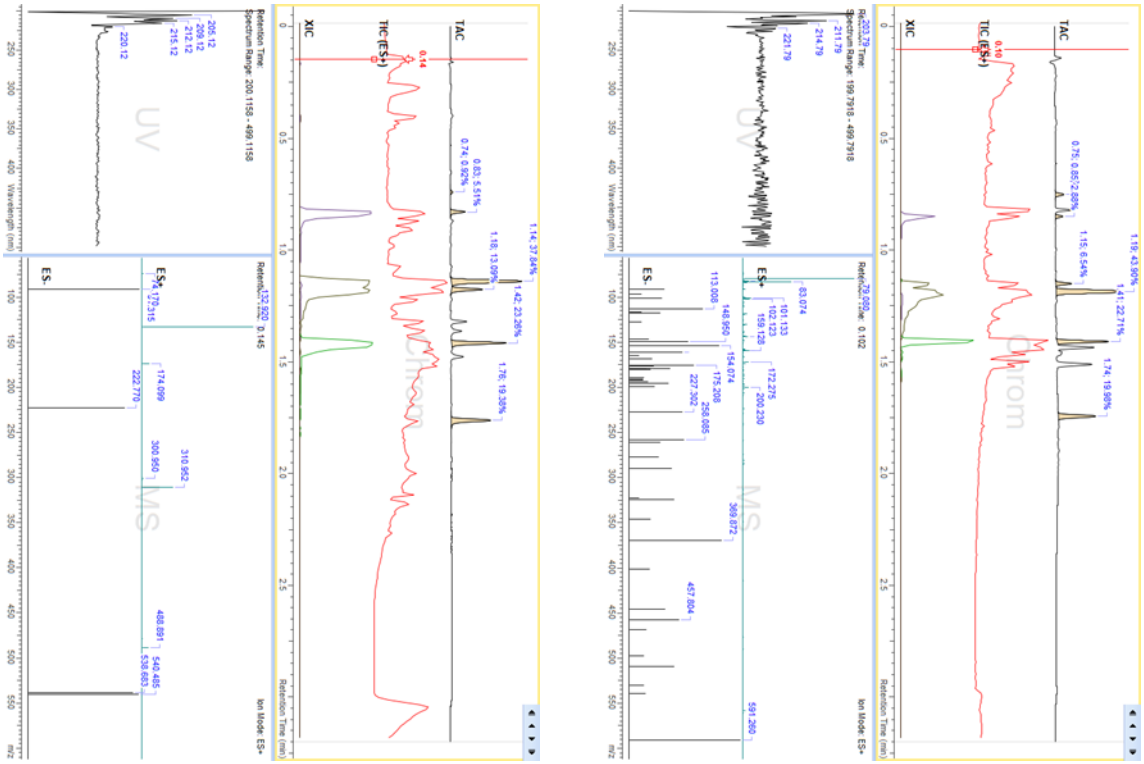


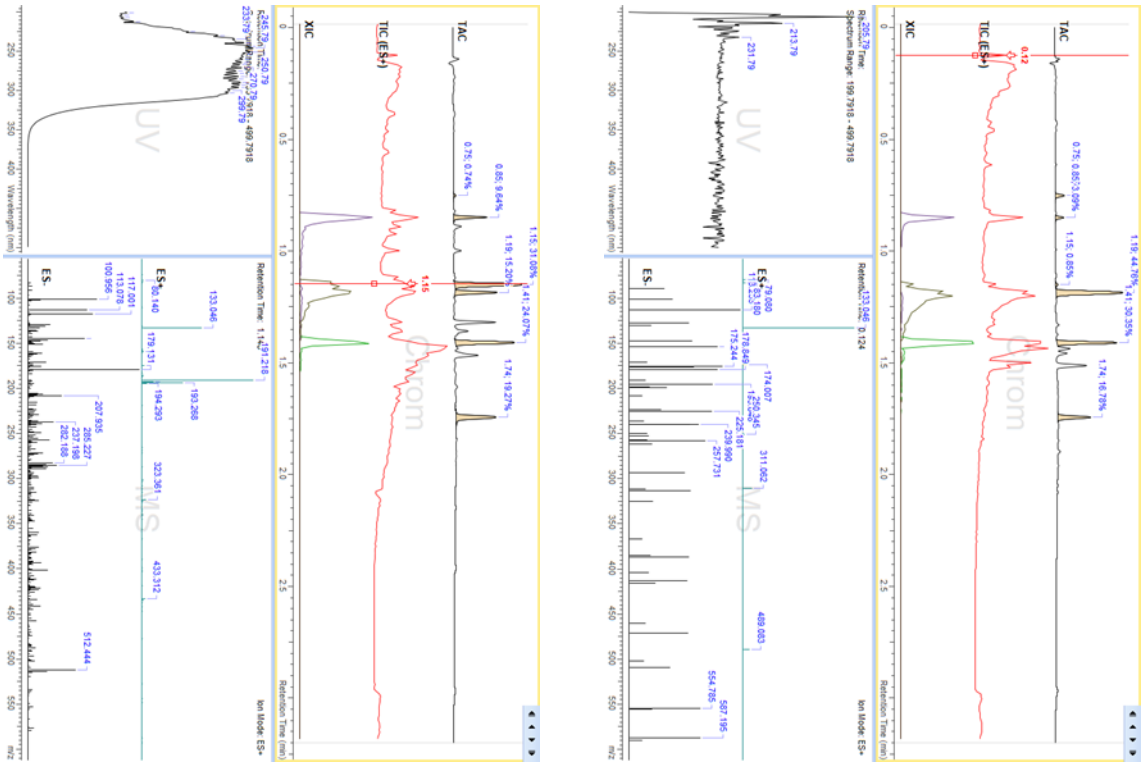
^{1}H NMR (400 MHz, DMSO-d_6) d ppm 7.56 - 7.63 (m, 3 H) 7.95 - 8.04 (m, 2 H) 8.29 (d, *J*=1.96 Hz, 1 H) 9.69 (d, *J*=1.96 Hz, 1 H)

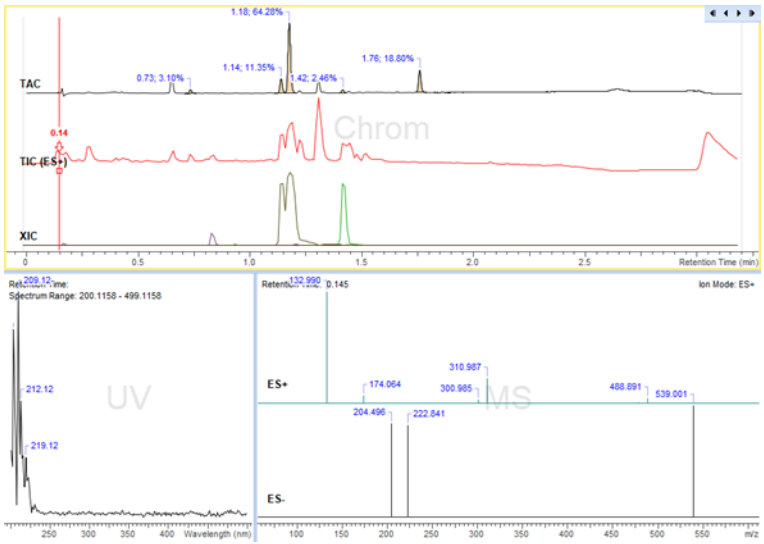


23

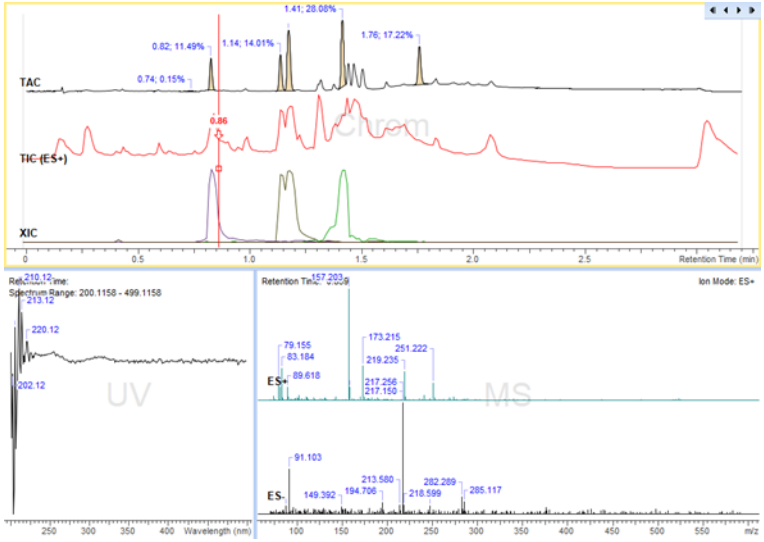


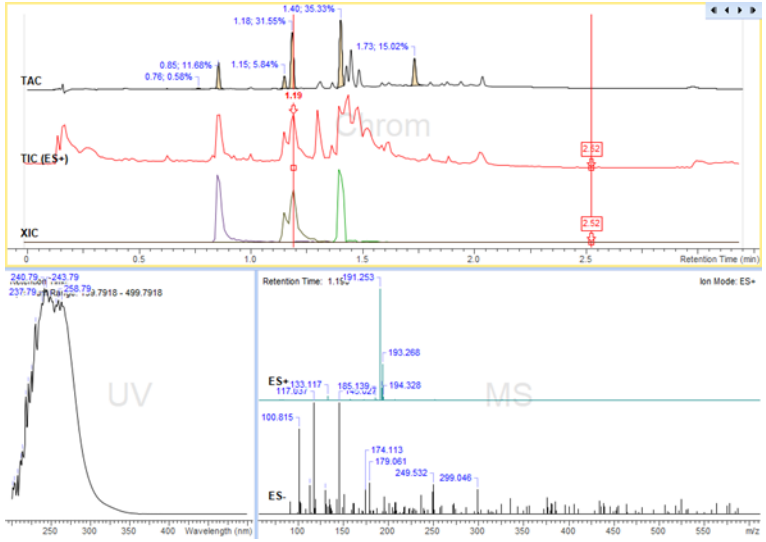


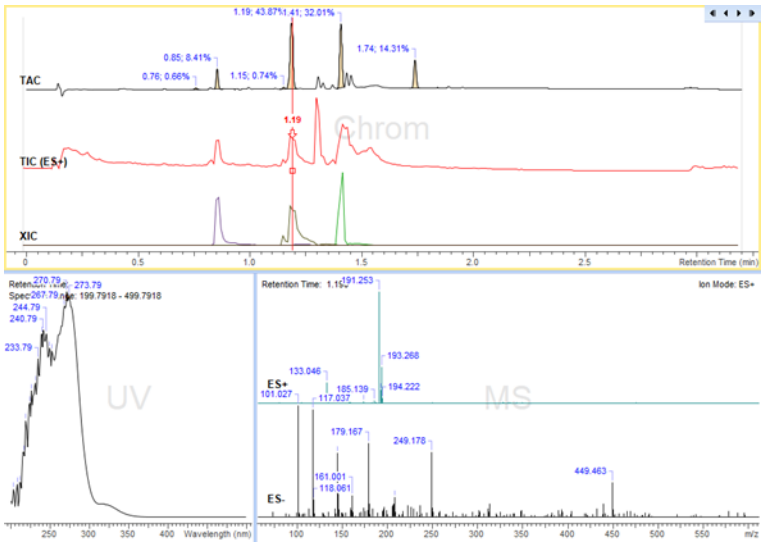


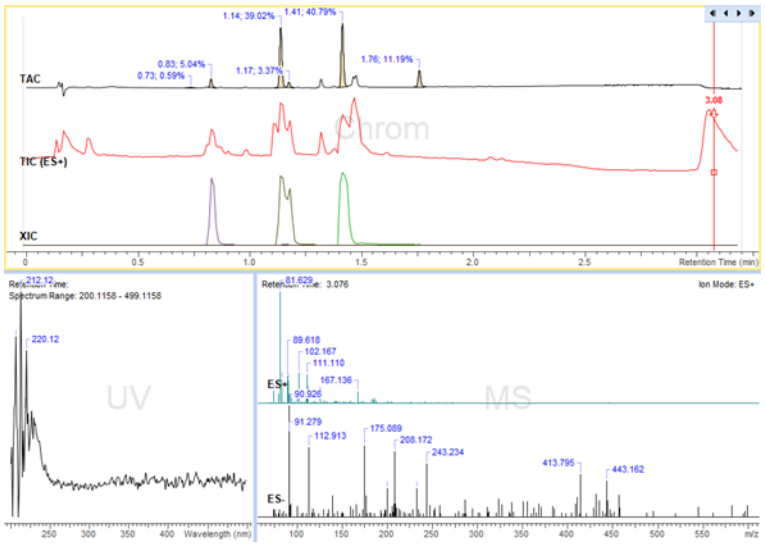

TAC
TIC (ES+)
XIC
Chrom
UV
MS
ES+
ES-
Ion Mode: ES+

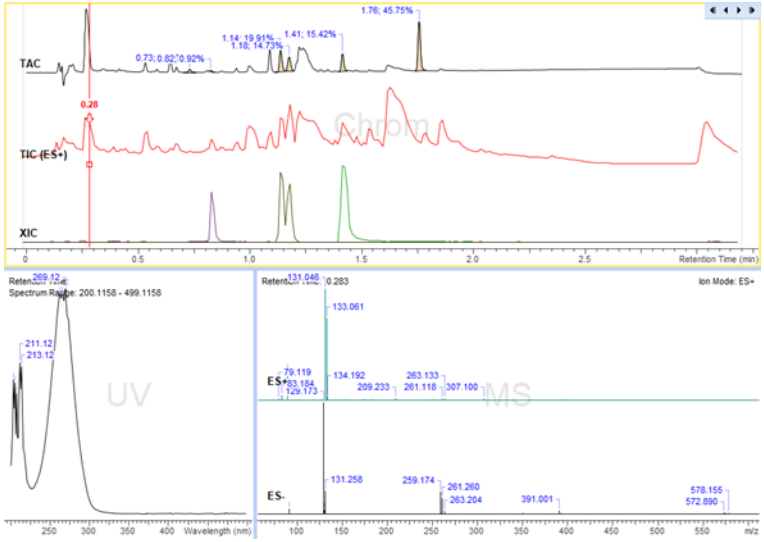

TAC
TIC (ES+)
XIC
Chrom
UV
MS
ES+
ES-
Ion Mode: ES+

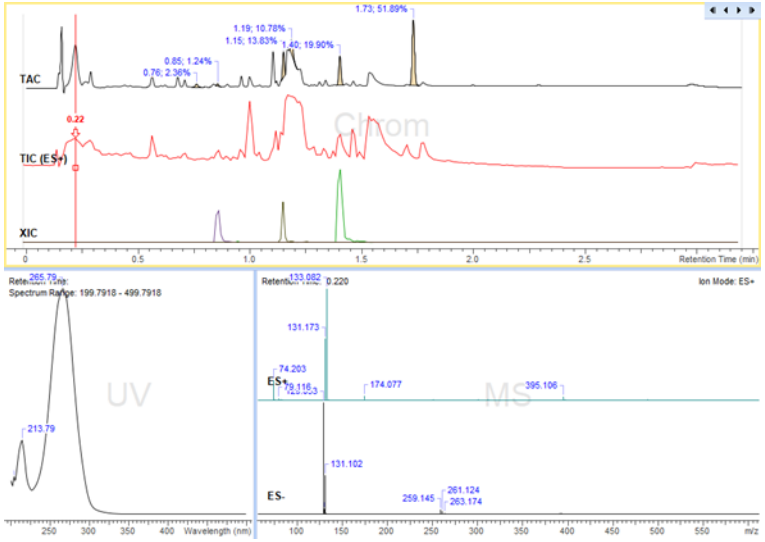

TAC
TIC (ES+)
XIC
Chrom
UV
MS
ES+
ES-
Ion Mode: ES+

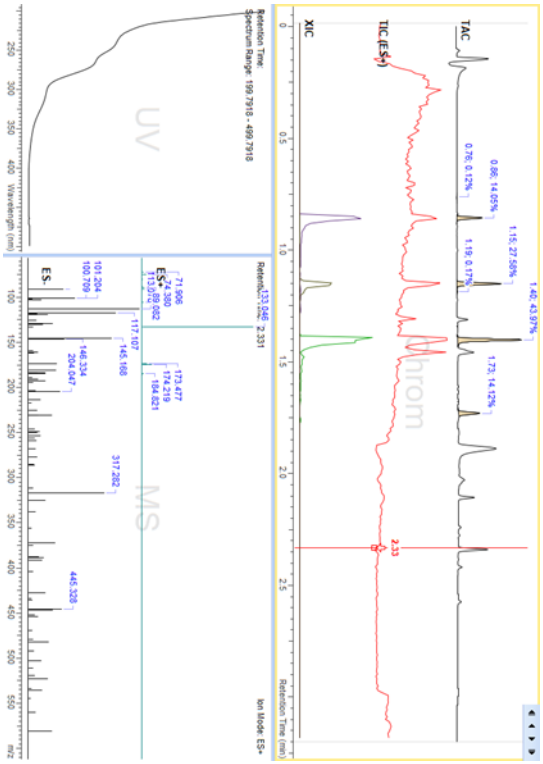

TAC
TIC (ES+)
XIC
Chrom
UV
MS
ES+
ES-
Ion Mode: ES+

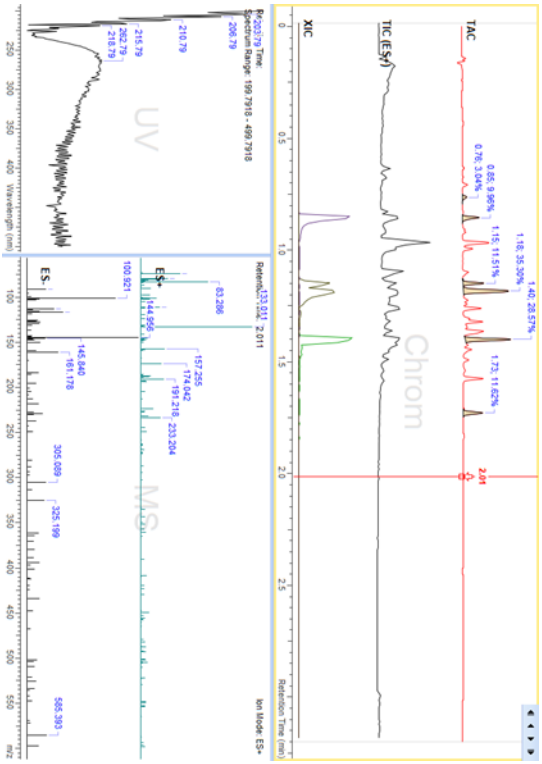

TAC
TIC (ES+)
XIC
Retention Time (min)
UV
MS
ES+
ES-
Ion Mode: ES+

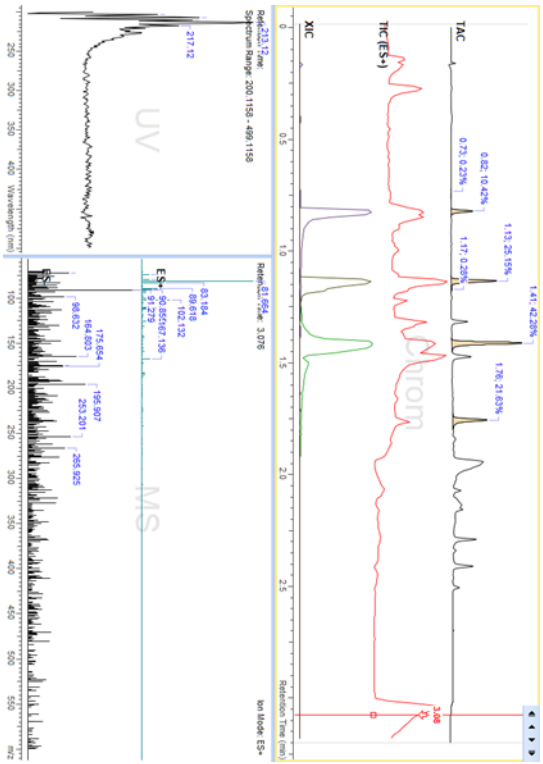

TAC
TIC (ES+)
XIC
Retention Time (min)
UV
MS
ES+
ES-
Ion Mode: ES+

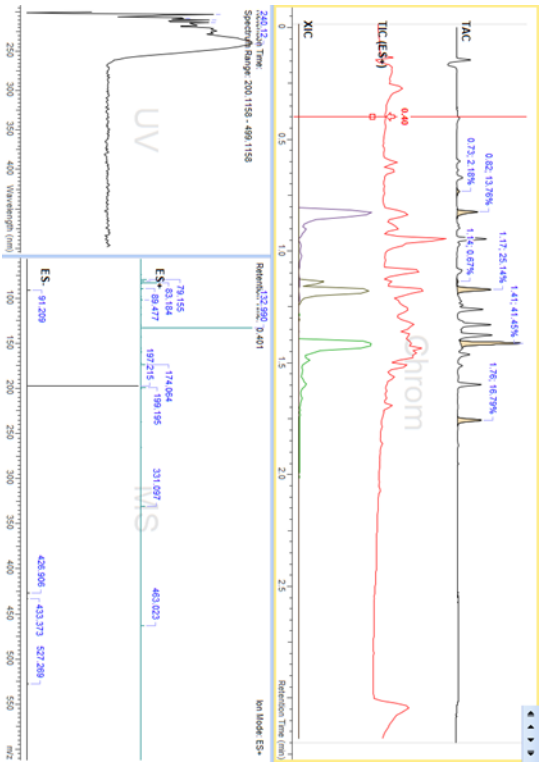

TAC
TIC (ES+)
XIC
Retention Time (min)
UV
MS
ES+
ES-
Ion Mode: ES+

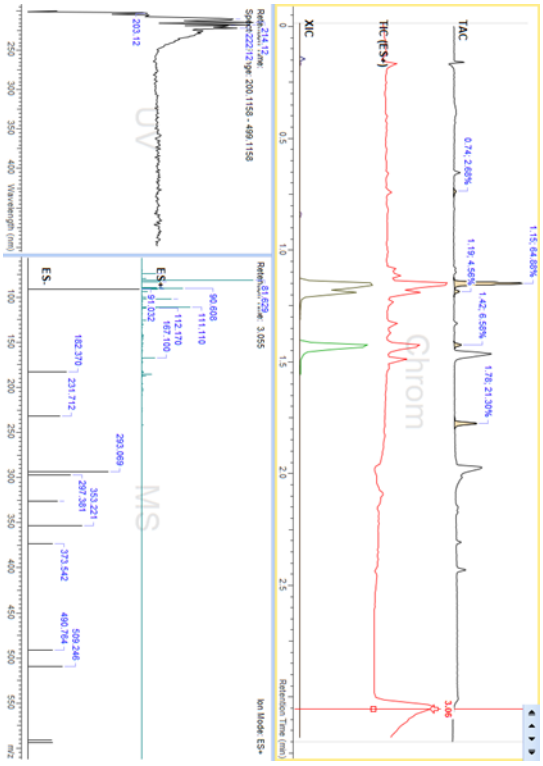

TAC
TIC (ES+)
XIC
Retention Time (min)
UV
MS
ES+
ES-
Ion Mode: ES+

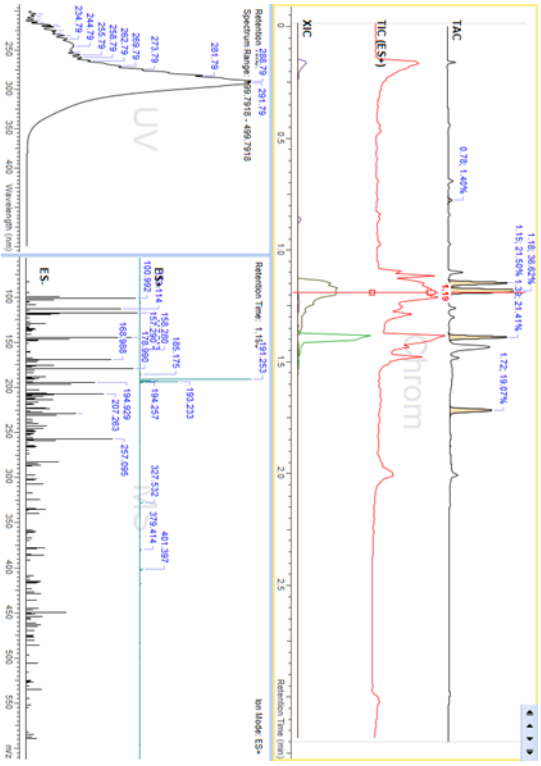

TAC
TIC (ES+)
XIC
Chrom
MS
UV
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z
Ion Mode: ES+

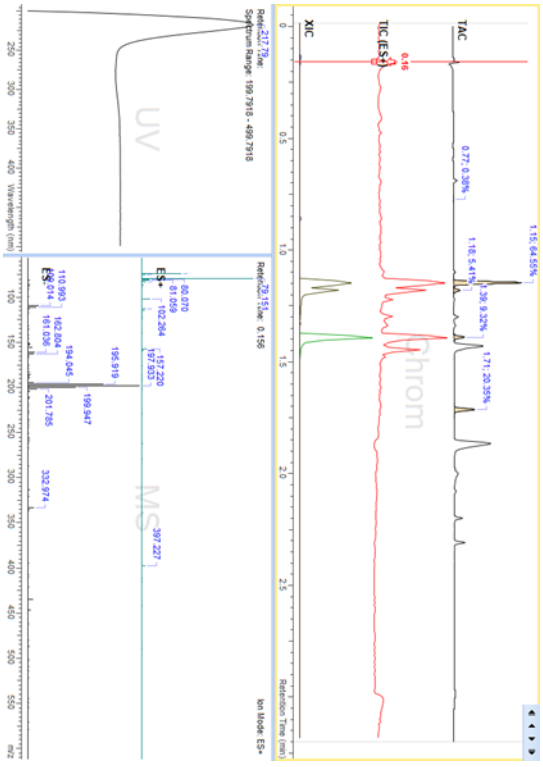

TAC
TIC (ES+)
XIC
Chrom
MS
UV
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z
Ion Mode: ES+

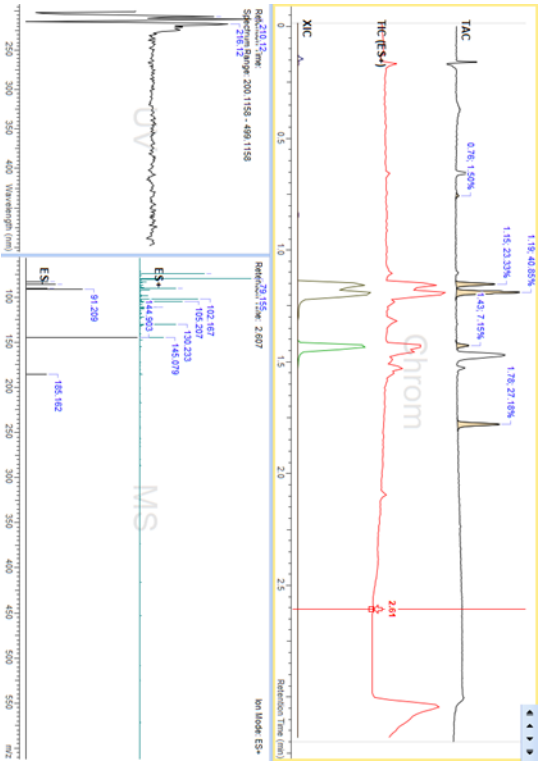

TAC
TIC (ES+)
XIC
Chrom
MS
UV
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z
Ion Mode: ES+

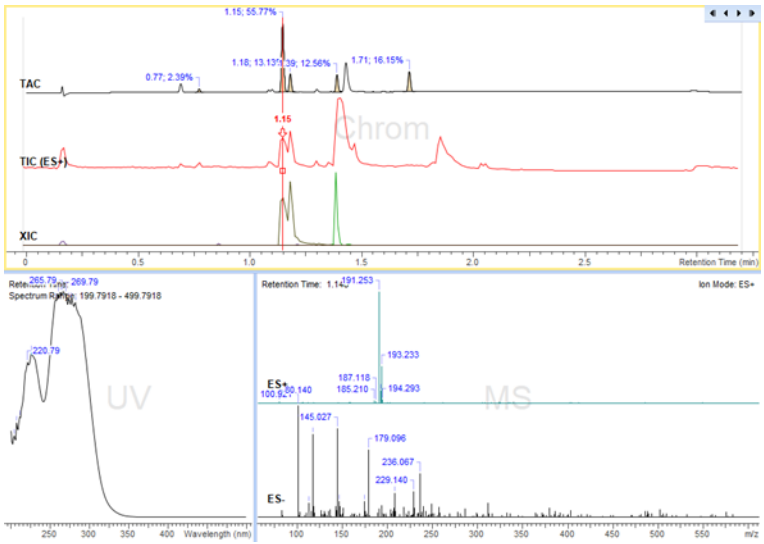

TAC
TIC (ES+)
XIC
Chrom
MS
UV
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z
Ion Mode: ES+

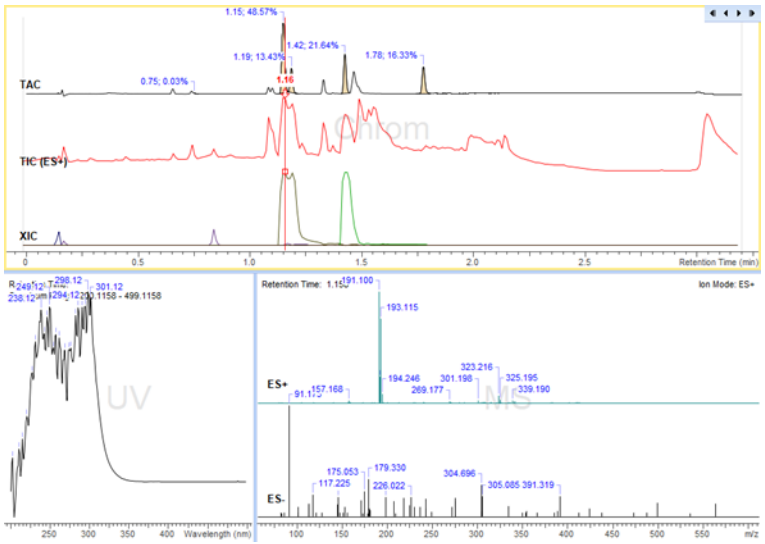


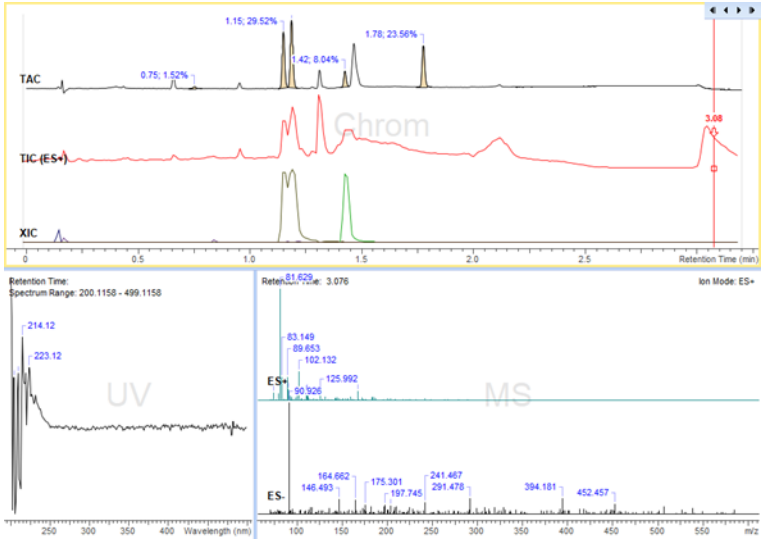


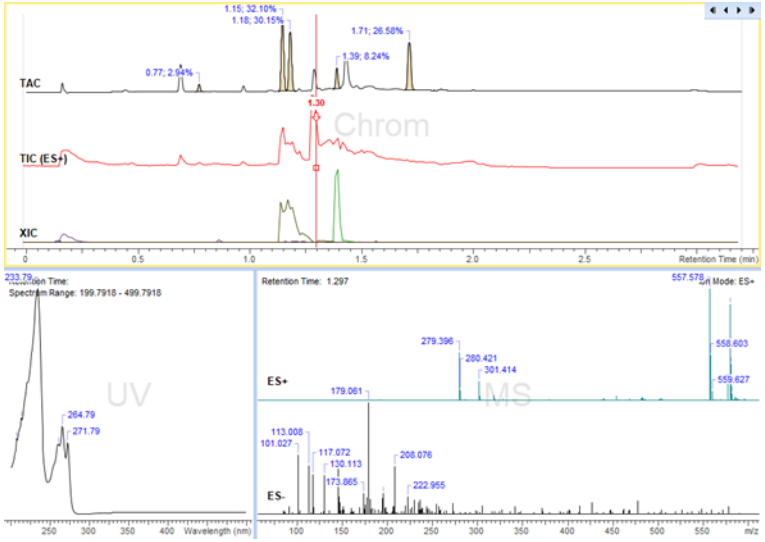


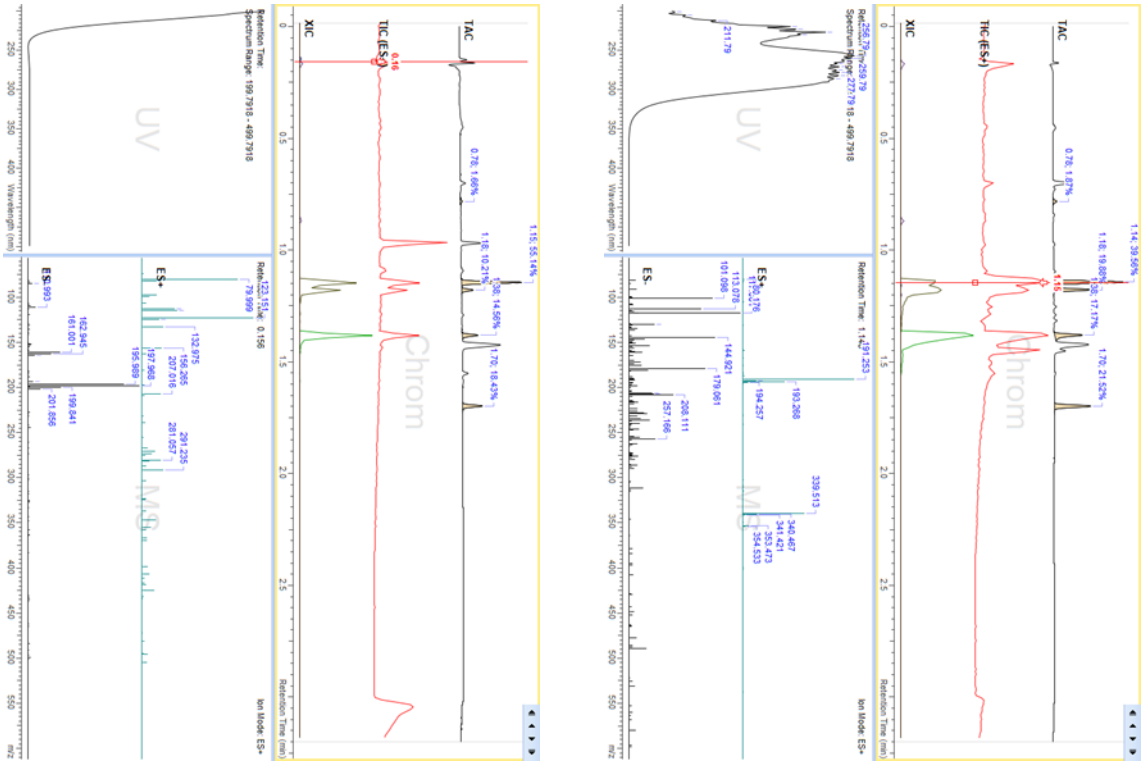


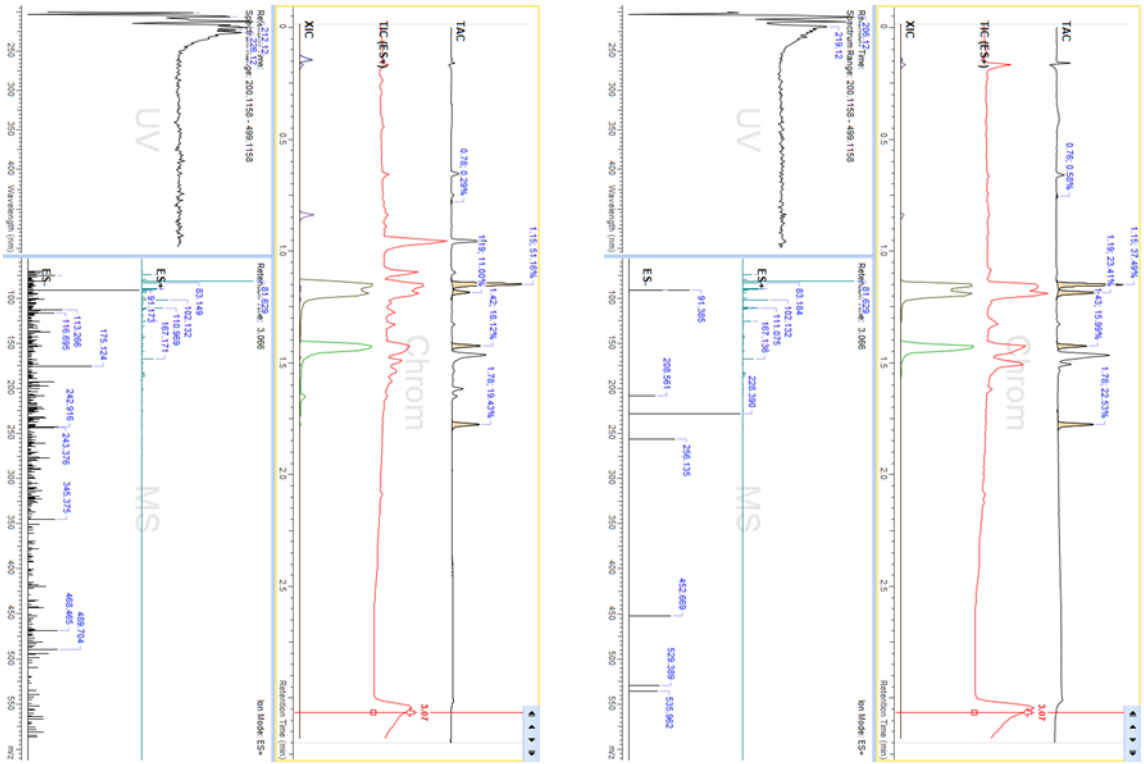

TAC
TIC (ES+)
XIC
Retention Time (min)
Wavelength (nm)
m/z
ES+
ES-
UV
MS

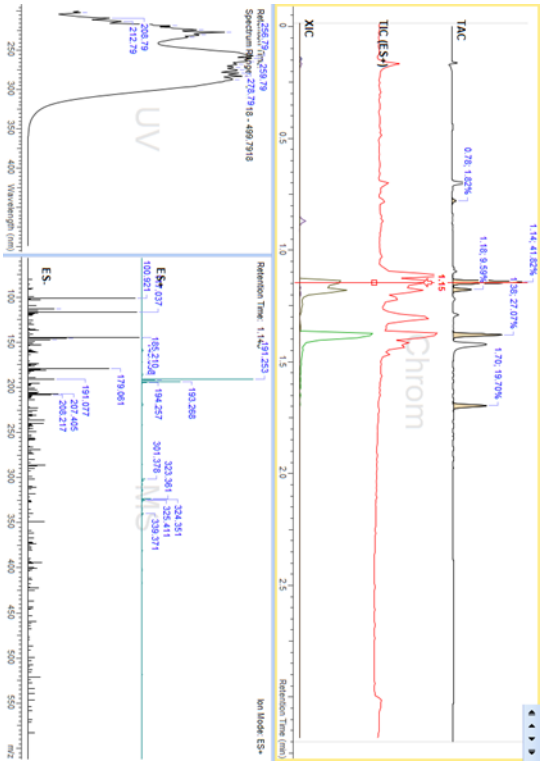

TAC
TIC (ES+)
XIC
Retention Time (min)
Wavelength (nm)
m/z
ES+
ES-
UV
MS

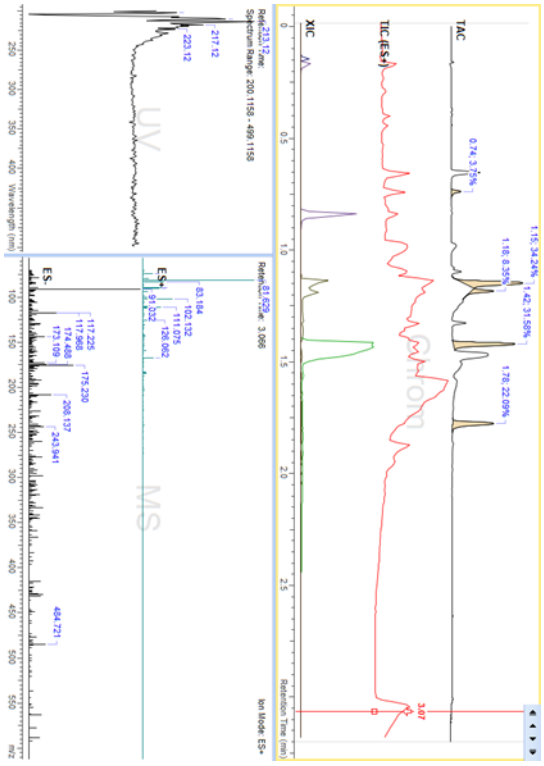

TAC
TIC (ES+)
XIC
Retention Time (min)
Wavelength (nm)
m/z
ES+
ES-
UV
MS

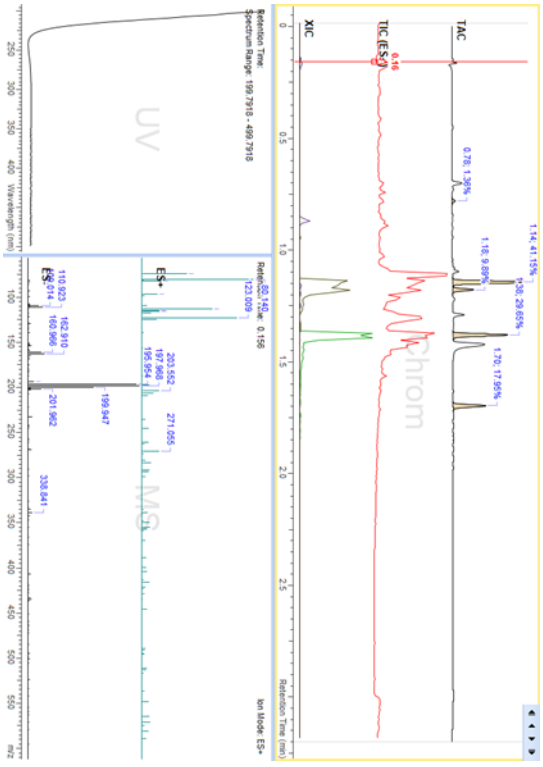

TAC
TIC (ES+)
XIC
Retention Time (min)
Wavelength (nm)
m/z
ES+
ES-
UV
MS

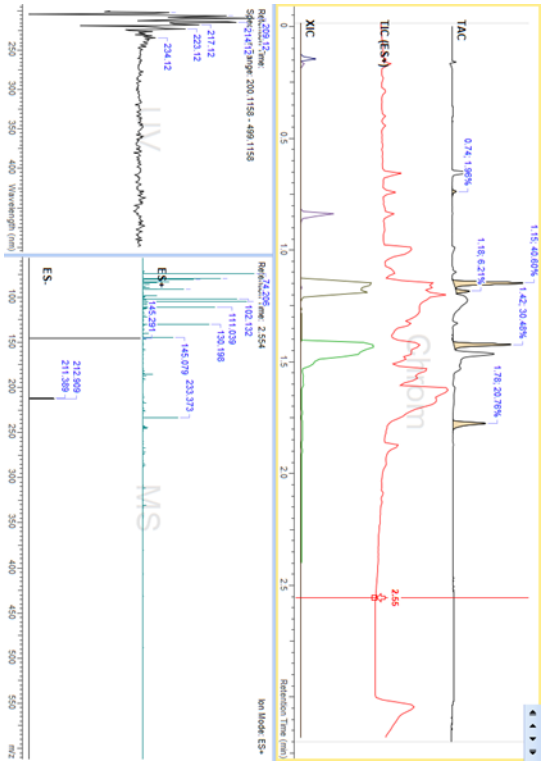

TAC
TIC (ES+)
XIC
UV
MS
Chrom

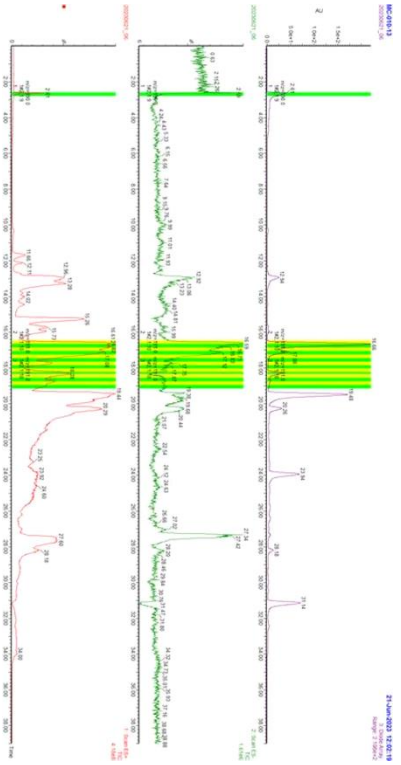

TAC
TIC (ES+)
XIC
UV
MS
Chrom

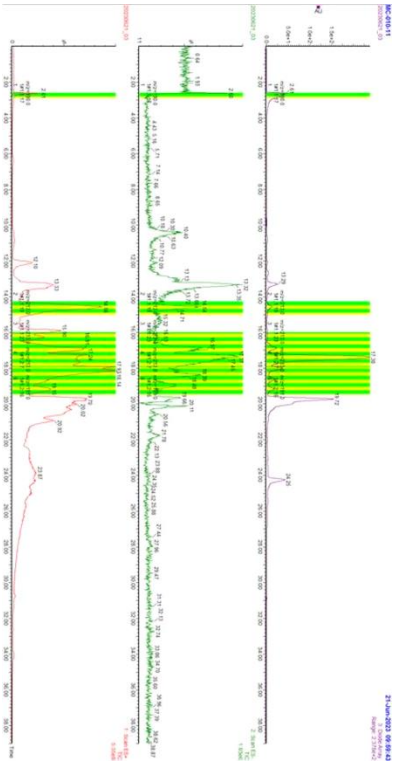

TAC
TIC (ES+)
XIC
UV
MS
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z

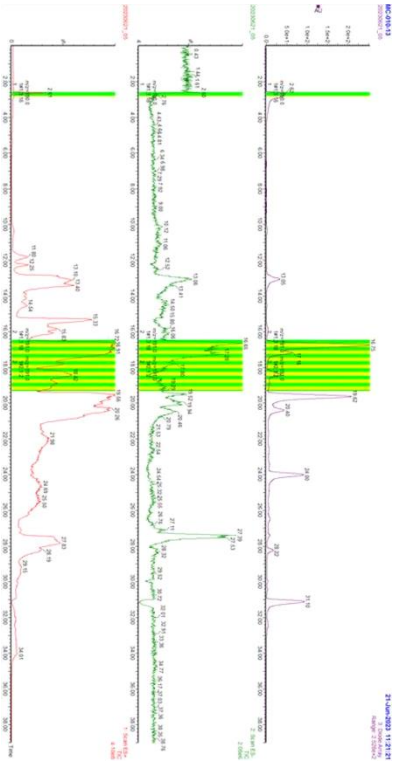

TAC
TIC (ES+)
XIC
UV
MS
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z

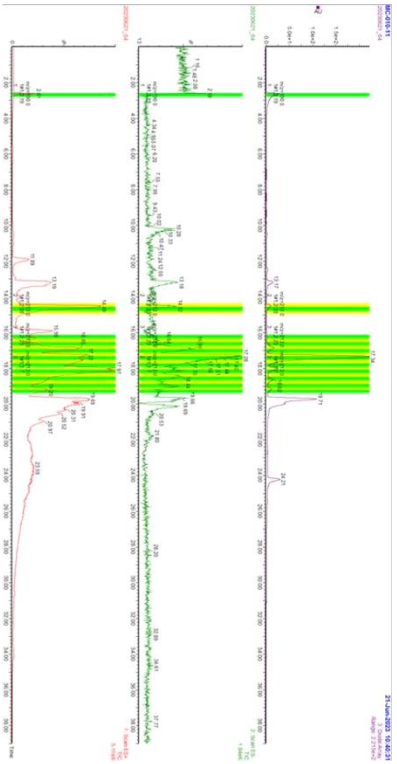

TAC
TIC (ES+)
XIC
UV
MS
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z


TAC
TIC (ES+)
XIC
UV
MS
ES+
ES-
Retention Time (min)
Wavelength (nm)
m/z









References

1. Norman, J. P. & Neufeldt, S. R. The Road Less Traveled: Unconventional Site Selectivity in Palladium-Catalyzed Cross-Couplings of Dihalogenated *N*-Heteroarenes. ACS Catalysis. 2022, vol. 12:19. P. 12014–12026. Available at: doi: 10.1021/acscatal.2c03743.

2. Norman, J. P., et al. Unconventional Site Selectivity in Palladium-Catalyzed Cross-Couplings of Dichloroheteroarenes under Ligand-Controlled and Ligand-Free Systems. The Journal of Organic Chemistry. 2022, vol. 87:11. P. 7414–7421. Available at: doi: 10.1021/acs.joc.2c00665.

3. Scott, N. W. J., et al. A Dichotomy in Cross-Coupling Site Selectivity in a Dihalogenated Heteroarene: Influence of Mononuclear Pd, Pd Clusters, and Pd Nanoparticles — the Case for Exploiting Pd Catalyst Speciation. Journal of the American Chemical Society. 2021, vol. 143:25. P. 9682–9693.Available at: doi: 10.1021/jacs.1c05294.

4. Dai, X., et al. Ligand-Dependent Site-Selective Suzuki Cross-Coupling of 3,5-Dichloropyridazines. The Journal of Organic Chemistry. 2013, vol. 78:15. P. 7758–7763. Available at: doi: 10.1021/jo401096u.

5. Keylor, M. H., et al. Inverting Conventional Chemoselectivity in Pd-Catalyzed Amine Arylations with Multiply Halogenated Pyridines. Journal of the American Chemical Society. 2017, vol. 139:31. P. 10613–10616. Available at: doi: 10.1021/jacs.7b05409.

6. Palani, V. & Perea, M. A. & Sarpong, R. Site-Selective Cross-Coupling of Polyhalogenated Arenes and Heteroarenes with Identical Halogen Groups. Chemical Reviews. 2021, vol. 122:11. 2021. P. 10126–10169. Available at: doi: 10.1021/acs.chemrev.1c00513.

7. Niemeyer, Z. L., et al. Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. Nature Chemistry. 2016, vol. 8:6. P. 610–617. Available at: doi: 10.1038/nchem.2501.

8. Santiago, C. B. & Guo, J. Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. Chemical Science. 2018, vol. 9:9. P. 2398–2412. Available at: doi: 10.1039/c7sc04679k.

9. Dotson, J. J., et al. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. Journal of the American Chemical Society. 2022, vol. 145:1. P. 110–121. Available at: doi: 10.1021/jacs.2c08513.

10. Panteleev, J. & Gao, H. & Jia, L. Recent applications of machine learning in medicinal chemistry. Bioorganic & Medicinal Chemistry Letters. 2018, vol. 28:17. P. 2807–2815. Available at: doi: 10.1016/j.bmcl.2018.06.046.

11. D'Alterio, M. C., et al. Mechanistic Aspects of the Palladium-Catalyzed Suzuki-Miyaura Cross-Coupling Reaction. Chemistry – A European Journal. 2021, vol. 27:54. P. 13481–13493. Available at: doi: 10.1002/chem.202101880.

12. Johansson Seechurn, C. C. C., et al. Palladium-Catalyzed Cross-Coupling: A Historical Contextual Perspective to the 2010 Nobel Prize. Angewandte Chemie International Edition. 2012, vol. 51:21. P. 5062–5085. Available at: doi: 10.1002/anie.201107017.

13. Farhang, M., et al. A retrospective-prospective review of Suzuki–Miyaura reaction: From cross-coupling reaction to pharmaceutical industry applications. Polyhedron. 2022, vol. 227. P. 16124. Available at: doi: 10.1016/j.poly.2022.116124.

14. García-Melchor, M., et al. Computational Perspective on Pd-Catalyzed C–C Cross-Coupling Reaction Mechanisms. Accounts of Chemical Research. 2013, vol. 46:11. P. 2626–2634. Available at: doi: 10.1021/ar400080r.

15. Casado, A. L. & Espinet, P. On the Configuration Resulting from Oxidative Addition of RX to $Pd(PPh_3)_4$ and the Mechanism of the cis-to-trans Isomerization of $[PdRX(PPh_3)_2]$ Complexes (R = Aryl, X = Halide). Organometallics. 1998, vol. 17:5. P. 954–959. Available at: doi: 10.1021/om9709502.

16. Fairlamb, I. J. S. Regioselective (site-selective) functionalisation of unsaturated halogenated nitrogen, oxygen and sulfur heterocycles by Pd-catalysed cross-couplings and direct arylation processes. Chemical Society Reviews. 2007, vol. 36:7. P. 1036. Available at: doi: 10.1039/b611177g.

17. Garcia, Y., et al. Theoretical Bond Dissociation Energies of Halo-Heterocycles: Trends and Relationships to Regioselectivity in Palladium-Catalyzed Cross-Coupling Reactions. Journal of the American Chemical Society. 2009, vol. 131:18. P. 6632–6639. Available at: doi: 10.1021/ja9004927.

18. Miyaura, N. & Suzuki, A. Palladium-Catalyzed Cross-Coupling Reactions of Organoboron Compounds. Chemical Reviews. 1995, vol. 95:7. P. 2457–2483. Available at: doi: 10.1021/cr00039a007.

19. Pudasaini, B. & Janesko, B. G. Computational Investigation of Selectivity in Suzuki–Miyaura Coupling of Secondary Alkyl Boranes. Organometallics. 2011, vol. 30:17. P. 4564–4571. Available at: doi: 10.1021/om200305r.

20. Fu, F., et al. A Robust and Efficient Pd_3 Cluster Catalyst for the Suzuki Reaction and Its Odd Mechanism. ACS Catalysis. 2017, vol. 7:3. P. 1860–1867. Available at: doi: 10.1021/acscatal.6b02527.

21. Rossi, R. & Bellina, F. & Lessi, M. Selective Palladium-Catalyzed Suzuki–Miyaura Reactions of Polyhalogenated Heteroarenes. Advanced Synthesis & Catalysis. 2012, vol. 354:7. P. 1181–1255. Available at: doi: 10.1002/adsc.201100942.

22. Almond-Thynne, J., et al. Site-selective Suzuki–Miyaura coupling of heteroaryl halides – understanding the trends for pharmaceutically important classes. Chemical Science. 2017, vol. 8:1. P. 40–62. Available at: doi: 10.1039/c6sc02118b.

23. Legault, C. Y., et al. Origin of Regioselectivity in Palladium-Catalyzed Cross-Coupling Reactions of Polyhalogenated Heterocycles. Journal of the American Chemical Society. 2007, vol. 129:42. P. 12664–12665. Available at: doi: 10.1021/ja075785o.

24. Schoenebeck, F. & Houk, K. N. Ligand-Controlled Regioselectivity in Palladium-Catalyzed Cross Coupling Reactions. Journal of the American Chemical Society. 2010, vol. 132:8. P. 2496–2497. Available at: doi: 10.1021/ja9077528.

25. Lu, J., et al. A reactivity model for oxidative addition to palladium enables quantitative predictions for catalytic cross-coupling reactions. Chemical Science. 2022, vol. 13:12. P. 3477–3488. Available at: doi: 10.1039/d2sc00174h.

26. Reeves, E. K., et al. Solvent Effects on the Selectivity of Palladium-Catalyzed Suzuki-Miyaura Couplings. Israel Journal of Chemistry. 2019, vol. 60: 3–4. P. 406–409. Available at: doi: 10.1002/ijch.201900082.

27. Scott, N. W. J., et al. The ubiquitous cross-coupling catalyst system '$Pd(OAc)_2$'/$2PPh_3$ forms a unique dinuclear PdI complex: an important entry point into catalytically competent cyclic Pd_3 clusters. Chemical Science. 2019, vol. 10:34. P. 7898–7906. Available at: doi: 10.1039/c9sc01847f.

28. Willans, C. E., et al. Cyclotriveratrylene-tethered trinuclear palladium(ii)–NHC complexes; reversal of site selectivity in Suzuki–Miyaura reactions. Dalton Transactions. 2019, vol. 48:39. P. 14687–14695. Available at: doi: 10.1039/c9dt03400e.

29. Martin, R. & Buchwald, S. L. Palladium-Catalyzed Suzuki—Miyaura Cross-Coupling Reactions Employing Dialkylbiaryl Phosphine Ligands. Accounts of Chemical Research. 2008, vol. 41:11. P. 1461–1473. Available at: doi: 10.1021/ar800036s.

30. Yang, W. & Wang, Y. & Corte, J. R. Efficient Synthesis of 2-Aryl-6-chloronicotinamides via PXPd2-Catalyzed Regioselective Suzuki Coupling. Organic Letters. 2003, vol. 5:17. P. 3131–3134. Available at: doi: 10.1021/ol035188g.

31. Manabe, K. & Yamaguchi, M. Catalyst-Controlled Site-Selectivity Switching in Pd-Catalyzed Cross-Coupling of Dihaloarenes. Catalysts. 2014, vol. 4:3. P. 307–320. Available at: doi: 10.3390/catal4030307.

32. Strotman, N. A., et al. Catalyst-Controlled Regioselective Suzuki Couplings at Both Positions of Dihaloimidazoles, Dihalooxazoles, and Dihalothiazoles. The Journal of Organic Chemistry. 2010, vol. 75:5. P. 1733–1739. Available at: doi: 10.1021/jo100148x.

33. Yang, M., et al. Palladium-Catalyzed C-4 Selective Coupling of 2,4-Dichloropyridines and Synthesis of Pyridine-Based Dyes for Live-Cell Imaging. The Journal of Organic Chemistry. 2020, vol. 85:10. P. 6498–6508. Available at: doi: 10.1021/acs.joc.0c00449.

34. Norman, J. P. & Larson, N. G. & Neufeldt, S. R. Different Oxidative Addition Mechanisms for 12- and 14-Electron Palladium(0) Explain Ligand-Controlled Divergent Site Selectivity. ACS Catalysis. 2022, vol. 12:15. P. 8822–8828. Available at: doi: 10.1021/acscatal.2c01698.

35. Williams, W. L., et al. The Evolution of Data-Driven Modeling in Organic Chemistry. ACS Central Science. 2021, vol. 7:10. P. 1622–1637. Available at: doi: 10.1021/acscentsci.1c00535.

36. Houk, K. N. & Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. Accounts of Chemical Research. 2017, vol. 50:3. P. 539–543. Available at: doi: 10.1021/acs.accounts.6b00532.

37. Taft, R. W. Jr. Linear Free Energy Relationships from Rates of Esterification and Hydrolysis of Aliphatic and Ortho-substituted Benzoate Esters. Journal of the American Chemical Society. 1952, vol. 74:11. P. 2729–2732. Available at: doi: 10.1021/ja01131a010.

38. Mater, A. C. & Coote, M. L. Deep Learning in Chemistry. Journal of Chemical Information and Modeling. 2019, vol. 59:6. P. 2545–2559. Available at: doi: 10.1021/acs.jcim.9b00266.

39. Sandfort, F., et al. A Structure-Based Platform for Predicting Chemical Reactivity. Chem. 2020, vol. 6:6. P. 1379–1390. Available at: doi: 10.1016/j.chempr.2020.02.017.

40. Lo, Y. C., et al. Machine learning in chemoinformatics and drug discovery. Drug Discovery Today. 2018, vol. 23:8. P. 1538–1546. Available at: doi: 10.1016/j.drudis.2018.05.010.

41. Yang, X., et al. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chemical Reviews. 2019, vol. 119:18. P. 10520–10594. Available at: doi: 10.1021/acs.chemrev.8b00728.

42. Chen, H., et al. The rise of deep learning in drug discovery. Drug Discovery Today. 2018, vol. 23:6. P. 1241–1250. Available at: doi: 10.1016/j.drudis.2018.01.039.

43. de Almeida, A. F. & Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. Nature Reviews Chemistry. 2019, vol. 3:10. P. 589–604. Available at: doi: 10.1038/s41570-019-0124-0.

44. Crawford, J. M., et al. Impact of Phosphine Featurization Methods in Process Development. Organic Process Research & Development. 2022, vol. 26:4. P. 1115–1123. Available at: doi: 10.1021/acs.oprd.1c00357.

45. Shi, Y., et al. Automated Experimentation Powers Data Science in Chemistry. Accounts of Chemical Research. 2021, vol. 54:3. P. 546–555. Available at: doi: 10.1021/acs.accounts.0c00736.

46. Selekman, J. A., et al. High-Throughput Automation in Chemical Process Development. Annual Review of Chemical and Biomolecular Engineering. 2017, vol. 8:1. P. 525–547. Available at: doi: 10.1146/annurev-chembioeng-060816-101411.

47. Sigman, M. S., et al. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. Accounts of Chemical Research. 2016, vol. 49:6. P. 1292–1301. Available at: doi: 10.1021/acs.accounts.6b00194.

48. Gensch, T., et al. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. Journal of the American Chemical Society. 2022, vol. 144:3. P. 1205–1217. Available at: doi: 10.1021/jacs.1c09718.

49. Bannwarth, C., et al. Extended tight-binding quantum chemistry methods. WIREs Computational Molecular Science. 2020, Vol. 11. Available at: doi: 10.1002/wcms.1493.

50. Grimme, S. & Bannwarth, C. & Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). Journal of Chemical Theory and Computation. 2017, vol. 13:5. P. 1989–2009. Available at: doi: 10.1021/acs.jctc.7b00118.

51. Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. Journal of Chemical Theory and Computation. 2019, vol. 15:5. P. 2847–2862. Available at: doi: 10.1021/acs.jctc.9b00143.

52. Bannwarth, C. & Ehlert, S. & Grimme, S. GFN2-xTB — An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. Journal of Chemical Theory and Computation. 2019, vol. 15:3. P. 1652–1671. Available at: doi: 10.1021/acs.jctc.8b01176.

53. Yap, C. W. PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 2010, vol. 32:7. P. 1466–1474. Available at: doi: 10.1002/jcc.21707.

54. Domingo, L. & Ríos-Gutiérrez, M. & Pérez, P. Applications of the Conceptual Density Functional Theory Indices to Organic Chemistry Reactivity. Molecules. 2016, vol. 21:6. P. 48. Available at: doi: 10.3390/molecules21060748.

55. Durand, D. J. & Fey, N. Computational Ligand Descriptors for Catalyst Design. Chemical Reviews. 2019, vol. 119:11. P. 6561–6594. Available at: doi: 10.1021/acs.chemrev.8b00588.

56. Brethomé, A. V. & Fletcher, S. P. & Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. ACS Catalysis. 2019, vol. 9:3. P. 2313–2323. Available at: doi: 10.1021/acscatal.8b04043.

57. Hoque, A. & Sunoj, R. B. Deep learning for enantioselectivity predictions in catalytic asymmetric β-C–H bond activation reactions. Digital Discovery. 2022, vol. 1:6. P. 926–940. Available at: doi: 10.1039/d2dd00084a.

58. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. Vol. 374:2065. 2016. P. 0150202. Available at: doi: 10.1098/rsta.2015.0202.

59. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021, vol. 2. Available at: doi: 10.1007/s42979-021-00592-x.

60. Freund, Y. & Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 1997, vol. 55:1. P. 119–139. Available at: doi: 10.1006/jcss.1997.1504.

61. Cervantes, J., et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 2020, vol. 408. P. 189–215. Available at: doi: 10.1016/j.neucom.2019.10.118.

62. Gardner, M. W. & Dorling, S. R. Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences. Atmospheric Environment. 1998, vol. 32: 14–15. P. 2627–2636. Available at: doi: 10.1016/S1352-2310(97)00447-0.

63. Mitchell, J. B. O. Machine learning methods in chemoinformatics. WIREs Computational Molecular Science. 2014, vol. 4:5. P. 468–481. Available at: doi: 10.1002/wcms.1183.

64. Ebi, T., et al. Design of Experimental Conditions with Machine Learning for Collaborative Organic Synthesis Reactions Using Transition-Metal Catalysts. ACS Omega. 2021, vol. 6:41. P. 27578–27586. Available at: doi: 10.1021/acsomega.1c04826.

65. Carpenter, K. A., et al. Deep learning and virtual drug screening. Future Medicinal Chemistry. 2018, vol. 10:21. P. 2557–2567. Available at: doi: 10.4155/fmc-2018-0314.

66. Ahneman, D. T., et al. Predicting reaction performance in C–N cross-coupling using machine learning. Science. 2018, vol. 360:6385. P. 186–190. Available at: doi: 10.1126/science.aar5169.

67. Collins, K. D. & Glorius, F. Intermolecular Reaction Screening as a Tool for Reaction Evaluation. Accounts of Chemical Research. 2015, vol. 48:3. P. 619–627. Available at: doi: 10.1021/ar500434f.

68. Rinehart, N. I., et al. A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. Science. 2023, vol. 381:6661. P. 965–972. Available at: doi: 10.1126/science.adg2114.

69. Angello, N. H., et al. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. Science. 2022, vol. 378:6618. P. 399–405. Available at: doi: 10.1126/science.adc8743.

70. Meyer, B., et al. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. Chemical Science. 2018, Vol. 9:35. P. 7069–7077. Available at: doi: 10.1039/C8SC01949E.

71. Mahjour, B. & Hoffstadt, J. & Cernak, T. Designing Chemical Reaction Arrays Using Phactor and ChatGPT. Organic Process Research & Development. 2023, vol. 27:8. P. 1510–1516. Available at: doi: 10.1021/acs.oprd.3c00186.

72. Beker, W., et al. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. Journal of the American Chemical Society. 2022, vol. 144:11. P. 4819–4827. Available at: doi: 10.1021/jacs.1c12005.

73. Gao, K., et al. Generative Network Complex for the Automated Generation of Drug-like Molecules. Journal of Chemical Information and Modeling. 2020, vol. 60:12. P. 5682–5698. Available at: doi: 10.1021/acs.jcim.0c00599.

74. Segler, M. H. S. & Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018, vol. 555:7698. P. 604–610. Available at: doi: 10.1038/nature25978.

75. Landrum, G. A., et al. RDKit: Open-source cheminformatics. https://www.rdkit.org

76. Nandy, A., et al. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. Chemical Reviews. 2021, vol. 121:16. P. 9927–10000. Available at: doi: 10.1021/acs.chemrev.1c00347.

77. Lempinen, A. Computational catalyst modeling and ligand performance classification for the Suzuki–Miyaura reaction. Master's thesis. University of Helsinki, Faculty of Science, Chemistry. Helsinki. 2023. 100 p.

78. Pracht, P. & Bohle, F. & Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. Physical Chemistry Chemical Physics. 2020, vol. 22:14. P. 7169–7192. Available at: doi: 10.1039/c9cp06869d.

79. De Jesus Silva, J., et al. Development and Molecular Understanding of a Pd-Catalyzed Cyanation of Aryl Boronic Acids Enabled by High-Throughput Experimentation and Data Analysis. Helvetica Chimica Acta. 2021, vol. 104. P. e 12. Available at: doi: 10.1002/hlca.202100200.

80. Zhao, S., et al. Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. Science. 2018, vol. 362:6415. P. 670–674. Available at: doi: 10.1126/science.aat2299.

81. Ziegler, T. & Tschinke, V. & Ursenbach, C. Thermal stability and kinetic lability of the metal carbonyl bond. A theoretical study on $M(CO)_6$ (M = chromium, molybdenum, tungsten), $M(CO)_5$ (M = iron, ruthenium, osmium), and $M(CO)_4$ (M = nickel, palladium, platinum) [Internet]. Vol. 109, Journal of the American Chemical Society. American Chemical Society (ACS); 1987. p. 4825–37. Available at: doi: 10.1021/ja00250a013.

82. Tolman, C. A. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. Chemical Reviews. 1977, vol. 77:3. P. 313–348. Available at: doi: 10.1021/cr60307a002.

83. Calderazzo, F. Carbonyl Complexes of the Transition Metals. Encyclopedia of Inorganic and Bioinorganic Chemistry. Wiley; 2005. Available at: doi: 10.1002/9781119951438.eibc0035.

84. Newman-Stonebraker, S. H., et al. Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. Science. 2021, vol. 374:6565. P. 301–308. Available at: doi: 10.1126/science.abj4213.

85. Jorner, K. MORFEUS. https://digital-chemistry-laboratory.github.io/morfeus/

86. Pedregosa, F., et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12. 2011. P. 2825–2830.

87. Harper, K. C. & Sigman, M. S. Three-Dimensional Correlation of Steric and Electronic Free Energy Relationships Guides Asymmetric Propargylation. Science. 2011, vol. 333:6051. P. 1875–1878. Available at: doi: 10.1126/science.1206997.

88. Schmidt, A. F. & Kurokhtina, A. A. & Larina, E. V. Role of a base in Suzuki-Miyaura reaction. Russian Journal of General Chemistry. 2011, vol. 81:7. P. 1573–1574. Available at: doi: 10.1134/S1070363211070334.

89. Amar, Y., et al. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. Chemical Science. 2019, vol. 10:27. P. 6697–6706. Available at: doi: 10.1039/c9sc01844a.

90. Ehlert, S., et al. Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. Journal of Chemical Theory and Computation. 2021, vol. 17:7. P. 4250–4261. Available at: doi: 10.1021/acs.jctc.1c00471.

91. Firsan, S. J. & Sivakumar, V. & Colacot, T. J. Emerging Trends in Cross-Coupling: Twelve-Electron-Based $L_1Pd(0)$ Catalysts, Their Mechanism of Action, and Selected Applications. Chemical Reviews. 2022, vol. 122:23. P. 16983–17027. Available at: doi: 10.1021/acs.chemrev.2c00204.

92. Vidossich, P. & Ujaque, G. & Lledós, A. Palladium monophosphine Pd(PPh3): is it really accessible in solution? In Chem. Commun. 2014, vol. 50:6. P. 661–663. Available at: doi: 10.1039/c3cc47404f.

93. Polshettiwar, V., et al. Suzuki–Miyaura Cross-Coupling Reactions in Aqueous Media: Green and Sustainable Syntheses of Biaryls. ChemSusChem. 2010, vol. 3:5. P. 502–522. Available at: doi: 10.1002/cssc.200900221.

94. Ackermann, M., et al. Electronic Properties of Furyl Substituents at Phosphorus and Their Influence on 31P NMR Chemical Shifts. Journal of the American Chemical Society. 2006, vol. 128:26. P. 8434–8440. Available at: doi: 10.1021/ja057085u.

95. Sinha, V. & Laan, J. J. & Pidko, E. A. Accurate and rapid prediction of pKa of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach. Physical Chemistry Chemical Physics. 2021, vol. 23:4. P. 2557–2567. Available at: doi: 10.1039/d0cp05281g.

96. Skoraczyński, G., et al. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? 2017, vol. 7. Scientific Reports. Springer Science and Business Media LLC. Available at: doi: 10.1038/s41598-017-02303-0.

97. Jablonka, K. M., et al. Leveraging large language models for predictive chemistry. Nature Machine Intelligence. 2024, vol. 6:2. P. 161–169. Available at: doi: 10.1038/s42256-023-00788-1.

98. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018, vol. 361:6400. P. 360–365. Available at: doi: 10.1126/science.aat2663.