



University of  
**Nottingham**

UK | CHINA | MALAYSIA

# **Large-Scale Computation and Prediction of Sulfinate-Mediated C-H Functionalisation Regioselectivity**

Thesis submitted to the University of Nottingham for the degree of  
Doctor of Philosophy

October, 2024

Peter Walton

14278442

Supervisors: Professor Jonathan Hirst, Professor Ross Denton

# Contents

Abstract .....	v
Acknowledgements .....	vi
List of Figures .....	vii
List of Abbreviations .....	x
Chapter 1 - Introduction .....	1
1.1 Machine Learning in Chemistry .....	1
1.1.1 Retrosynthesis and Synthetic Route Planning .....	2
1.1.2 Protein Folding .....	5
1.1.3 Catalysis .....	8
1.1.4 Metabolism Prediction .....	10
1.2 Machine Learning and Quantum Chemistry .....	11
1.2.1 Machine Learning and QM Methods .....	12
1.2.2 Molecular Property and Reaction Prediction .....	13
1.3 ML/QM for Reactivity .....	16
1.4 C-H Functionalisation .....	16
1.5 Project Goals .....	22
Chapter 2 - Background and Related Work .....	24
2.1 Quantum Chemistry .....	24

2.1.1	The Schrödinger Equation .....	24
2.1.2	Hartree-Fock Theory .....	29
2.1.3	Semi-empirical Methods .....	30
2.1.4	Density Functional Theory .....	31
2.1.5	Fukui Indices.....	36
2.1.6	Transition State and Frequency Calculation Procedure .....	38
2.1.7	Solvent Models and Other QM Calculation Techniques.....	40
2.2	Machine Learning.....	42
2.2.1	Classical QSAR Modelling Methods .....	42
2.2.2	Graph Neural Networks .....	50
2.3	Related Work in Prediction of C-H Functionalisation Regiochemistry .....	54
2.4	Application of Theory in this Work .....	57
Chapter 3 -	Predicting Regioselectivity .....	<b>Error! Bookmark not defined.</b>
3.1	Introduction and Methods.....	58
3.2	Results and Discussion .....	61
3.2.1	Atom-Centred Charges.....	61
3.2.2	Activation Energy .....	68
3.3	Activation Energy Calculations on other Drug-like Compounds.....	76
3.4	Conclusions .....	78

Chapter 4 -	Automation and Software Development.....	80
4.1	Local Calculation .....	80
4.1.1	Transition State Templates .....	82
4.1.2	Semi-empirical Calculations .....	84
4.2	HPC Calculation .....	87
4.2.1	Modification of AM1 transition state .....	88
4.2.2	Generation of HPC input file and HPC resource allocation .....	89
4.2.3	Job Submission and Monitoring.....	90
4.2.4	Resubmission and Data Collation .....	92
4.3	Graphical User Interface and Database Management .....	93
4.4	Conclusion .....	96
Chapter 5 -	Dataset Generation and Machine Learning.....	97
5.1	Dataset Generation .....	97
5.1.1	Dataset Curation .....	97
5.1.2	Large-Scale Calculations and Data Preprocessing.....	99
5.1.3	Traditional Modelling.....	101
5.2	ChemProp Modelling .....	104
5.2.1	Chemprop Regression Models .....	104
5.2.2	ChemProp Classification Models .....	108

Chapter 6 -	Conclusions and Outlook .....	111
Chapter 7 -	Appendix .....	115
7.1	General Experimental Information .....	115
7.2	Synthesis of Trifluoromethylated Literature Compounds .....	117
7.3	Activation Energy Calculation Data .....	130
Chapter 8 -	Bibliography .....	139

# Abstract

Sulfinate-mediated radical C–H functionalisation reactions are widely used for the modification and diversification of scaffolds in drug discovery. However, prediction of the regiochemistry in these reactions can be challenging. For a given substrate, there may be multiple sites of reaction, each with its own unique steric and electronic environment. Here we present Rega, an automated transition state searching program for the prediction of regioselectivity from inexpensive HF/6-31G\* activation energies. We show that in a set of 23 compounds, the regioselectivity is correctly identified in 22 cases (reactivity correctly identified for 65/68 potential sites of reaction). The easy-to-use and modular Rega workflow allows reaction exploration of multiple substrates simultaneously, enabling the generation of a synthetic dataset of 490 compounds consisting of 2780 sites with labelled reactivity for this reaction for use in machine learning models. Rega is designed to be readily extensible to other reaction systems and can be applied to many other reaction classes in which a radical intermediate is formed as the regiochemistry determining step. From the generation of this dataset, machine learning was applied to predict regioselectivity in both regression and classification tasks.

# Acknowledgements

I would like to thank my supervisors Prof. Jonathan Hirst and Prof. Ross Denton for their support and guidance throughout this work. My thanks also go to Dr Andrew Baxter for his ideas and encouragement and to Dr Kristaps Ermanis for his support in the development of the Rega platform. My thanks go to all of the Nottingham Computational Chemistry department and the Denton group for their support and encouragement.

# List of Figures

<b>Figure 1.1:</b> Overview of single and multi-step retrosynthesis tools, including goals, methods and evaluation metrics .....	5
<b>Figure 1.2</b> Simulation of heterogeneous catalysis from the atomic level to large-scale reactors .....	8
<b>Figure 1.3</b> Examples of C-H functionalisation with varying selectivity-determining processes.....	18
<b>Figure 1.4.</b> Oxidative C–H functionalisation reactions of heteroarenes using zinc and sodium sulfinates. ....	19
<b>Figure 1.5:</b> Mechanism for metal sulfinate-mediated C-H functionalisation .....	20
<b>Figure 1.6:</b> Free energy diagram for each step in zinc sulfinate-mediated C-H functionalisation computed at CPCM(water)-M06-2X/6-311+G(d,p)// B3LYP/6-31+G(d) level of theory .....	21
<b>Figure 2.1</b> Jacob’s ladder approach for the systematic improvement of DFT functionals.....	36
<b>Figure 2.2</b> The relationship between the dependent variable (y) and the independent variable (x).....	43
<b>Figure 2.3:</b> Random samples from (a) prior distribution of functions and (b) posterior distribution of functions in a one-dimensional example .....	48



<b>Figure 2.4:</b> Graph representation example of methylbenzene .....	51
<b>Figure 2.5:</b> Basic neural network architecture.....	52
<b>Figure 2.6:</b> ChemProp neural network architecture.....	54
<b>Figure 3.1:</b> Site selectivity is determined by activation energy. ....	60
<b>Figure 3.2</b> Set of 10 compounds used for the initial regioselectivity investigation. .	62
<b>Figure 3.3</b> Set of 16 drug-like compounds ( <i>a-p</i> ) used to evaluate Hirshfeld charge as a predictor of regioselectivity. ....	65
<b>Figure 3.4</b> Benchmark set of 18 compounds ( <i>a-r</i> ) used to evaluate Hirshfeld charge's regioselective predictions .....	67
<b>Figure 3.5</b> Activation energies (kcal mol <sup>-1</sup> ) for each site in the preliminary set of ten compounds.....	70
<b>Figure 3.6</b> Activation energies (kcal mol <sup>-1</sup> ) for each potential site of reaction in the benchmark set and the experimentally observed ratios of each product .....	72
<b>Figure 3.7</b> Experimental regioselectivities of compounds that previously disagreed with calculation .....	75
<b>Figure 3.8</b> Other drug-like compounds where activation energy was compared with experiment. ....	76
<b>Figure 4.1</b> A flow diagram outlining the full functionality of Rega.....	81
<b>Figure 4.2</b> Workflow for the semi-empirical calculation portion of Rega. ....	87
<b>Figure 4.3</b> HPC workflow of Rega. ....	88

<b>Figure 4.4</b> Graphical user interface for single compound calculation on Rega.....	95
<b>Figure 5.1</b> Funnel plot of the number of compounds remaining after each filtering step.....	98
<b>Figure 5.2</b> Example prediction vs calculation for gradient boosting (XGBoost) method.....	103
<b>Figure 5.3:</b> Calculated vs predicted activation energies for the ChemProp regression model using typical reaction SMILES. ....	107
<b>Figure 5.4</b> t-SNE plot of the clustered DrugBank dataset.....	110

# List of Abbreviations

ADCH	Atomic Dipole Corrected Hirshfeld charge
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
AM1	Austin Model 1
AUROC	Area Under the Receiver Operator Curve
B3LYP	Becke, 3-parameter, Lee-Yang-Parr
CASP	Critical Assessment of protein Structure Prediction
CCSD	Coupled Cluster Singles and Doubles
DFT	Density Functional Theory
DFTB	Density-Functional Tight-Binding
FFN	Feed-Forward Network
GGA	Generalised Gradient Approximation

GNN	Graph Neural Network
GP	Gaussian Processes
GUI	Graphical User Interface
HF	Hartree-Fock Theory
HOMO	Highest Occupied Molecular Orbital
HPC	High-Performance Compute
LDA	Local Density Approximation
LUMO	Lowest Unoccupied Molecular Orbital
M06-2x	Minnesota 06
MCTS	Monte Carlo Tree Search
mGGA	Meta-Generalised Gradient Approximation
ML	Machine Learning
MMFF94	Merck Molecular Force Field 94

MP2	Møller–Plesset 2
MRSA	Methicillin-Resistant <i>Staphylococcus Aureus</i> (MRSA)
MSA	Multiple Sequence Alignment
NDDO	Neglect of Differential Diatomic Overlap
NMR	Nuclear Magnetic Resonance
PES	Potential Energy Surface
PM7	Parameterised Method 7
QM	Quantum Mechanics
QSAR	Quantitative Structure-Activity Relationship
QTAIM	Quantum Theory of Atoms in Molecules
RF	Random Forests
RMSE	Root Mean Square Error
SCF	Self-Consistent Field

SMARTS	SMILES Arbitrary Target Specification
SMD	Solvation Model based on Density
SMILES	Simplified Molecular Input Line Entry System
TS	Transition State
UDP	Uridine Diphosphate

# Chapter 1 - Introduction

## 1.1 Machine Learning in Chemistry

Machine learning (ML) is having a profound impact in chemistry. From analytical and computational disciplines to the automation of full laboratories and scientific workflows, machine learning is changing how the modern chemist conducts their research and has the potential to dramatically accelerate breakthroughs in a wide range of areas.<sup>1</sup> Machine learning's ability to identify patterns in data allows chemists to gain new insights into their research in a less time and resource-intensive manner than traditional methods and allows for easier transferability to new areas of work. Some key areas of interest in the application of machine learning in chemistry, particularly of interest to drug discovery are reaction development, quantitative structure-activity relationship (QSAR) and absorption, distribution, metabolism, excretion and toxicity (ADMET) studies. Through the use of a variety of different depictions of compounds in a dataset known as descriptors, chemical structures are transformed to a machine-readable vector and are then fed into a machine learning algorithm and used to infer the inherent features of each compound that give rise to the output value being studied. This can be used to predict the output value of a compound that has not been studied before. This estimate can be used to inform decisions on the next steps in research, saving time and cost (i.e. avoiding synthesis

of a compound that is predicted to be inadequate for the project goal since the predicted result is given instantaneously).

In this section, we discuss in more detail some key applications of machine learning in chemistry and their impact on the area of drug discovery. Firstly, we will explore broader applications of machine learning in fields such as retrosynthesis, protein folding and metabolite prediction and their utility to the pharmaceutical industry. Next, we will discuss machine learning implementations in the context of computational chemistry through the generation of new ML-derived quantum chemistry methods, prediction of molecular properties. Lastly, we will discuss recent implementations of machine learning for the prediction of reaction outcomes and reactivity. This serves to give a broader understanding of the importance of machine learning in chemistry and its' impact on a wide array of fields.

### 1.1.1 Retrosynthesis and Synthetic Route Planning

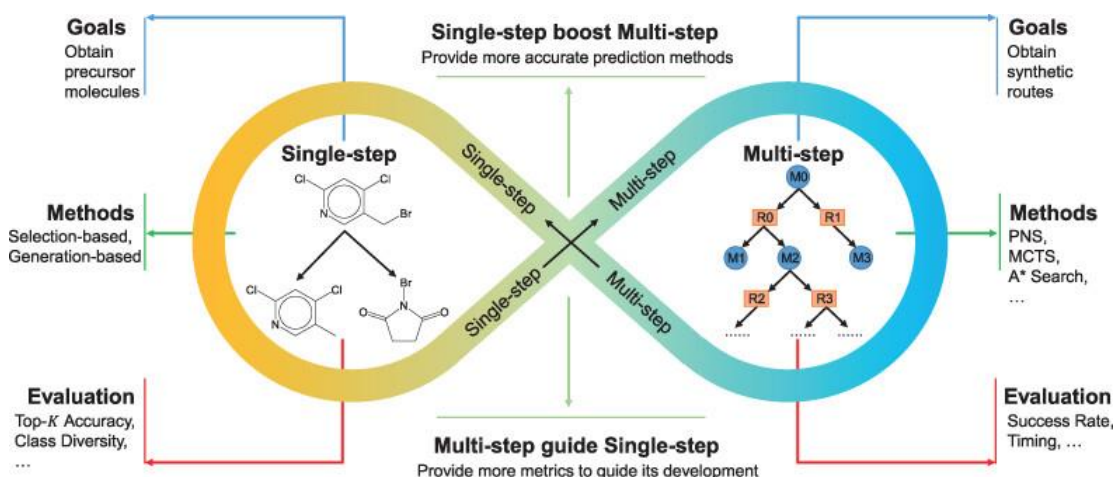
The cost of bringing a new therapeutic drug to market was estimated to be \$2.3 billion in 2023,<sup>2</sup> highlighting the need for efficiencies to be found in every step of the drug discovery process. One key area is the improvement in synthetic chemistry method development.<sup>3</sup> In the past, synthetic routes to new compounds were developed based on a trial-and-error search of appropriate reactions for pre-selected starting materials to give the desired product<sup>4</sup>. This approach lacks flexibility and often leads to the selection of a suboptimal route.<sup>5</sup> Retrosynthetic analysis serves to



combat this problem by decomposing the final molecule into several simpler precursor compounds<sup>6</sup> and defining the appropriate reactions to complete the conversion of reactant to product. With the ever-growing library of different reactions developed each year, it becomes impossible for an organic chemist to know every possible synthetic route to their desired compound, as it is thought that 10,000 different chemical transformations could be considered in each step of the synthesis.<sup>7</sup> Therefore, tools were developed to navigate this reaction landscape and suggest the most viable synthetic route to the user. While initial iterations of these retrosynthesis applications were based upon hard-coded rules of reaction types,<sup>8</sup> this hindered the real-world application. Thus, deep learning was applied to the retrosynthetic problem,<sup>9–12</sup> which provided greater flexibility in the selection of reaction classes and gave a higher chance of these routes being effective in the real world. Retrosynthesis tools can be separated into two distinct classes. Single-step tools primarily focus on selecting the most appropriate precursors for a single transformation; multi-step tools attempt to decompose a more complex target compound into a full synthetic route comprising a multitude of different reactions (**Figure 1.1**). Data used to train these models are typically derived from the US patent office,<sup>13</sup> a large open-source database of over 1,000,000 different chemical reactions. These tools then evaluate the similarity of the user's input compound with each reaction in the database to determine the most suitable route to their desired compound. One key multi-step example was the 3N-MCTS tool,<sup>14</sup> which combined a Monte-Carlo tree search<sup>15</sup> with deep learning to provide a multi-step retrosynthesis tool capable of suggesting appropriate chemical transformations faster than any

previous attempts. A notable example of single-step retrosynthesis software is Zhang's work on the development of an evolutionary algorithm to optimise the search space and suggest routes 83.9% faster than other single-step Monte-Carlo methods.<sup>16</sup> Another promising retrosynthetic tool is SynPlanner, which utilises Monte-Carlo tree search and graph neural networks to generate a multi-step synthesis tool with features such as a value network to provide a measure of synthetic feasibility for each suggested route.<sup>17</sup> Lastly, the current state-of-the-art retrosynthesis tool is AiZynthFinder, an open-source tool using Monte-Carlo tree search alongside a neural network (see section 2.2.2) to suggest synthetic routes from a library of known reaction templates. One key differentiator here is the incorporation of precursor-pricing metrics to determine not only the most synthetically tractable route but also the most commercially viable.<sup>18</sup> Another exciting extension of ML in retrosynthesis is the implementation into computer-aided synthesis planning. In this work, retrosynthesis tools are used to offer a synthetic route to the desired compound as before. Once the route is confirmed by the user, a series of actions are defined for a robotic arm to carry out the synthesis directly without the need for human intervention. One key example of this work is ASKCOS,<sup>19</sup> where their retrosynthesis prediction algorithm was paired with a configurable flow chemistry apparatus. This newly developed workflow enables chemists to easily synthesise a compound of interest using chemically viable reactions without having to perform the direct synthesis themselves. Robotic synthesis allows for high levels of control over the reaction and greatly increases synthetic reproducibility.

While the databases used to train these retrosynthetic tools are very large, they do not account for the entirety of chemical space and so depending on the region of chemical space being studied, model applicability may be poor. Thus, more work is being done to generate models with larger datasets and with a greater understanding of the impact other functional groups may have on a given reaction's feasibility.



**Figure 1.1:** Overview of single and multi-step retrosynthesis tools, including goals, methods and evaluation metrics. Figure reproduced from reference.<sup>4</sup>

### 1.1.2 Protein Folding

One prominent application of machine learning in drug discovery is the prediction of protein folding. When developing a new therapeutic compound, the drug's structure is tailored to fit the receptor identified as playing a key role in the disease of interest. While great strides have been made to isolate and obtain protein conformational data on over 100,000 proteins<sup>20</sup> using a variety of experimental techniques,<sup>21–24</sup> this is only a small fraction of the billions of protein sequences known.<sup>25,26</sup> In order to

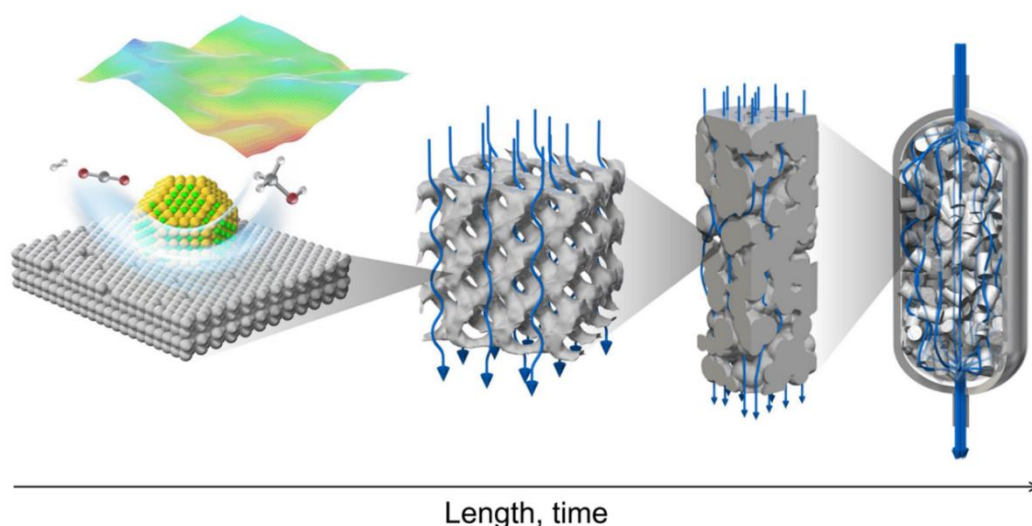
circumvent the lengthy process of isolating and obtaining an X-ray crystal structure for every protein of interest, an accurate estimation of the protein conformation through machine learning techniques was needed to accelerate drug discovery. The protein folding problem – the deduction of 3-dimensional protein conformation based solely on amino acid sequence has been a challenge in the field of bioinformatics for over 50 years.<sup>27,28</sup> The current completed benchmark used to evaluate the performance of protein folding tools is the 15th critical assessment of protein structure prediction (CASP15),<sup>29</sup> an unseen amino acid sequence is given where the model's output conformation is validated against experiment to determine the conformational accuracy. While other methods<sup>30–34</sup> of conformational prediction had some level of agreement with experiment, there was one model that provided near atom-level accuracy of conformational prediction, AlphaFold.<sup>35</sup> AlphaFold utilises a neural network with a unique understanding of biological and geometrical constraints of amino acids that enable an accuracy of 0.96 Å RMSE against the CASP15 benchmark, with the next best model only achieving an accuracy of 2.8 Å. This level of predictive performance has changed the landscape of life sciences, enabling huge developments in a variety of fields, from the modelling of antibiotic-resistant enzymes in bacteria;<sup>36</sup> to finding new treatments for Parkinson's disease through the identification and mapping of the mutations of the PINK1 protein, thought to be a leading factor for the development of early-onset Parkinson's.<sup>37–39</sup> The AlphaFold system takes the amino acid sequence as an input and constructs a multiple sequence alignment (MSA) to identify common sequence structures with known conformations from living organisms. This process enables

the identification of parts of the sequence where conformation is more likely to differ from one another. An initial structural template is built from proteins with a similar structure to the input sequence to determine the amino acid residues that are likely to be in contact with one another and feeds this alongside the multiple sequence alignment into a transformer model. The transformer model refines the multiple sequence alignment through comparison with the structural template and vice versa, in an iterative process until a specified number of cycles is reached. The final part of the pipeline takes these refined sequence alignment and structure representations to construct a three-dimensional model of the structure. This entire process is also iterative, where the refined MSA, template and predicted structure is fed back into the model to refine its structural prediction further.

Advancements in this field are continuing today, with the launch of AlphaFold3 substantially improving both protein structure prediction but also protein-ligand interaction behaviour.<sup>40</sup> This breakthrough in protein structure prediction has led to AlphaFold being used as a primary tool for protein structure analysis, highlighting the benefits of machine learning in science. This work has led to the award of the 2024 Nobel Prize in chemistry, specifically Demis Hassabis and John Jumper of DeepMind for the development of the AlphaFold program and to David Baker of the University of Washington for the design of proteins using AlphaFold to create steroid binding proteins with high affinity and selectivity,<sup>41</sup> the creation of self-assembling protein macrostructures<sup>42</sup> and the creation of protein switches and sensors<sup>43</sup> for the detection of the widely abused opioid fentanyl.<sup>44</sup>

### 1.1.3 Catalysis

Catalysts are essential for the development of new chemical processes in an efficient manner. Catalytic species serve to offer a new reaction pathway to reach a desired compound while avoiding the use of energy- and resource-intensive processes used prior. The shape of the catalyst plays a key role in the determination of the reaction's feasibility, and therefore catalyst design is of great importance in the advancement of these processes. Typically, heterogeneous catalysis optimisation requires the use of molecular dynamics simulations with long timescales to effectively model the interactions between catalyst and substrate (**Figure 1.2**).<sup>45</sup> This process of optimisation remains a formidable task, and as such, many novel catalysts and catalytic reactions are still discovered using a trial-and-error approach. This highlights the need for efficiencies to be found in the field of catalysis through the incorporation of machine learning.



**Figure 1.2** Simulation of heterogeneous catalysis from the atomic level to large-scale reactors. Figure reproduced from reference.<sup>45</sup>

Since in homogeneous catalysis, a catalyst's particular steric or electronic property governs the performance of a given material, initial work was done in the prediction of one of these materials' properties such as atomisation energies,<sup>46,47</sup> formation energies<sup>48,49</sup> and band gaps.<sup>50</sup> This approach however is not directly applicable to heterogeneous catalysis due to the complex nature of the interactions between catalyst and substrate not being dictated by a single property. Initial work has been done in the direct prediction of catalytic rates and selectivity using machine learning,<sup>51</sup> with Hattori *et al.* pioneering studies in the development of machine learning for catalyst design.<sup>52–54</sup> Design of novel heterogeneous catalysts using machine learning for the oxidative coupling of methane to new C<sub>2</sub> products including ethane and ethene has recently been achieved through understanding the relationship between catalyst physical properties and catalytic activity.<sup>55</sup> Another application of machine learning in catalysis is the prediction of acid catalyst activity for the promotion of a Friedel-Crafts reaction.<sup>56</sup> This was done through the learning of the relationship between physiochemical properties of different acid additives and reaction activity using Gaussian Processes Regression (see section 2.2.1) and predictions were verified through experiment.

ML has also been shown to be a powerful tool in photocatalysis. Using data mining, 540 cases of photocatalytic water-splitting were studied and enabled the generation of rules that govern the effectiveness of hydrogen generation in an energy-efficient manner.<sup>57</sup> This has great possible implications for the future of global CO<sub>2</sub> reduction through the development of hydrogen-based energy solutions, reducing reliance on fossil fuels. Lastly, the discovery of new CO<sub>2</sub> hydrogenation catalysts using ML-

derived activity predictions has the potential to further reduce climate change. Effective carbon capture and conversion to long-chain hydrocarbons both reduce the need for new fossil fuel mining and actively reduce atmospheric carbon dioxide, a key contributor to global warming. This work found that iron-based catalysts with potassium, zinc or copper additives promoted high conversion of CO<sub>2</sub> with high selectivity towards C<sub>5</sub>-C<sub>15</sub> hydrocarbons. The addition of TiO<sub>2</sub> was also essential for high selectivity against conversion to CH<sub>4</sub>, another key greenhouse gas. These advances in catalyst design using ML show the great impact of these techniques in not only reaction development but also the possibility of tackling some of the biggest issues facing the world today.

#### 1.1.4 Metabolism Prediction

Another important application of machine learning in drug discovery is the prediction of metabolic pathways. This process identifies potential sites of reaction for the compound of interest, informs the user if the compound is likely to be metabolised by a particular enzyme, and identifies the possible metabolites of the xenobiotic of interest. Doing so allows the user to understand whether their compound of interest is suitable for their desired application, and if potentially harmful by-products of metabolism are formed, then changes to the structure can be made early to avoid costly mistakes later in drug development. Successful models have been generated for the most likely xenobiotic metabolising enzyme found in the human body, Cytochrome P450,<sup>58-61</sup> as well as lesser-known enzymes which can also metabolise drug-like compounds, UDP-glucuronosyltransferases and flavin-containing



monooxygenases.<sup>62</sup> The modelling of the entire metabolic profile of xenobiotics is a continuing challenge, as the collection of biological data introduces a large amount of noise through varying conditions in the biological assays and different enzyme isoforms having varying activity towards each xenobiotic. As such, the true enzymatic activity towards these drug compounds can be difficult to measure and accurately predict using machine learning, since data quality is imperative for the understanding of the structure/function relationship. Nonetheless, recent advances in biological assay data acquisition and modelling have led to the development of commercially viable tools commonly used in the pharmaceutical industry. These models have had a profound impact on the workflow of the medicinal chemist, allowing more targeted optimisation of a lead compound in order to bring the final drug through to clinical trials in a shorter timescale.

## 1.2 Machine Learning and Quantum Chemistry

Quantum Chemistry has been profoundly impacted by the incorporation of machine learning. From the direct use of these techniques to generate new quantum mechanical calculation methods<sup>63</sup> to the prediction of properties and reaction pathways, machine learning has benefitted the quantum chemistry field greatly. Previously, if an individual desired to understand the intricacies of a particular compound, be it the mechanistic pathway that compound took to generate a given product; or the calculation of certain molecular properties; it required the

recruitment of knowledgeable professionals to run calculations and report their findings. Nowadays, the insights that machine learning methods have provided into the applications of quantum mechanics and the relationship between structure and experimental observations have enabled the generation of machine learning models that can dramatically improve the scientific discovery workflow. These models allow for more synergistic collaboration between QM and experimental fields and dramatically accelerate the rate at which breakthroughs can be found. In this section, we discuss some of the key implementations of machine learning in the field of quantum chemistry and its impact on the drug discovery landscape.

### 1.2.1 Machine Learning and QM Methods

In order to offer accurate predictions of experimental observation, appropriate quantum mechanical calculation methods must be selected. In each new instance of these investigations, a trade-off must be made between the level of accuracy and the computational cost of the calculation. Many rapidly calculated methods do not offer a level of accuracy suitable to be indicative of experiment, as a great number of approximations must be made which do not reflect the true nature of the interactions between subatomic particles taking place within a molecule. Therefore, more costly computational methods are often selected, at the expense of increased lead times. The use of machine learning in the development of new computational methods typically aims to offer the level of accuracy seen in more costly traditional methods in a fraction of the time.<sup>64–85</sup> Some key examples include the ability to use much longer timescales in molecular photodynamic simulations,<sup>69</sup> gaining a greater

understanding of the mechanistic behaviour of rapid photo-induced reactions; or a dramatically increased accuracy of simulations of the same timescale with the incorporation of quantum mechanical effects in machine-learned force fields.<sup>73</sup> Machine learning is highly dependent on the quality of the input data, making it ideal for application in quantum chemistry where simulations are highly controlled and repeatable. Depending on the dataset type used, ML can be trained on costly *ab-initio* methods and then predicted energies can be fed into lower cost simulations to provide additional accuracy.<sup>77</sup> Another possible implementation is improving density functional theory accuracy by augmenting exchange-correlation functionals such as B3LYP with additional high-order contributions to the energy from a neural network<sup>86</sup> or increasing efficiency by avoiding the explicit calculation of key components through the use of machine learning-derived predicted values.<sup>87,88</sup> It is worth noting the importance of the domain of applicability of these ML/QM methods. While machine learning can be extremely capable of interpolation, extrapolation ability is poor. Therefore, typical ML/QM publications involve the training of a model to solve a specific chemical problem rather than generating a general model aimed at broad applicability to a wide range of chemical systems.<sup>89–99</sup>

### 1.2.2 Molecular Property and Reaction Prediction

Another key area of efficiency improvements to the drug discovery process is the use of machine learning to predict molecular properties and reaction outcomes. These methods aim to reduce the cost of drug development through the reduction of the “trial-and-error” procedures traditionally used in drug design.<sup>100</sup> These tools allow

for rapid optimisation of synthetic route development and lead compound design by allowing medicinal chemists to test their hypotheses of different reaction pathways and scaffold substitutions without the need for wet lab experiments, which can be extremely time and resource-intensive. In these models, the relationship between the compound and the predicted property is learned through the use of chemical descriptors. These descriptors give the model an understanding of the full molecular environment as well as local atom-specific environments. The features of the environment that give rise to either a given molecular property or reaction outcome are learned and then used to predict new examples for the end-user. The underpinnings of machine learning and their relation to reactivity modelling are described later in this work (Chapter 2).

Some examples of molecular property prediction pertinent to drug discovery are the prediction of pharmacokinetic parameters.<sup>101</sup> These properties are traditionally measured in a biological assay which can be a costly and difficult process to generate accurate data. Also, the inherent variability of biological systems introduces noise into the measurement leading to uncertainty on the validity of the result. Machine learning implementation on tasks such as these allows for the estimation of these properties in an expedient manner and enables rapid decision-making on the direction of the research. Examples such as the prediction of half-life, clearance, volume of distribution and fraction bound/unbound were shown to be effective for cephalosporins and a variety of structurally diverse antihistamines.<sup>102,103</sup>

Seminal work in the prediction of reaction outcome prediction was carried out by Doyle *et al.*,<sup>104</sup> who utilised random forest modelling to map the relationship between calculated molecular/vibrational descriptors and reaction yield for the Buchwald-Hartwig C-N cross-coupling reactions. This work was limited to the reaction between singly substituted aryl halides and 4-methylaniline, but the promising predictive performance of these models warranted further expansion from other groups. Yield prediction of Buchwald-Hartwig amination reactions was extended to new amine derivatives with great success.<sup>105,106</sup> The use of deep learning with graph neural networks<sup>105</sup> saw the generation of industrially viable models that allow the user to understand whether their planned reaction will work as intended before having to invest the time and resources performing the experiment. One other example of this extension was conducted by the Hirst group,<sup>107</sup> who utilised support vector regression modelling rather than random forests to predict reaction yield on the same combinatorial dataset used in their work. They saw a large increase in  $R^2$  and a reduction in RMSE over methods previously attempted. This shows that the continuing advancements in the development of new machine learning techniques can improve model performance and generalisability over time in the field of reaction outcome prediction.

## 1.3 ML/QM for Reactivity

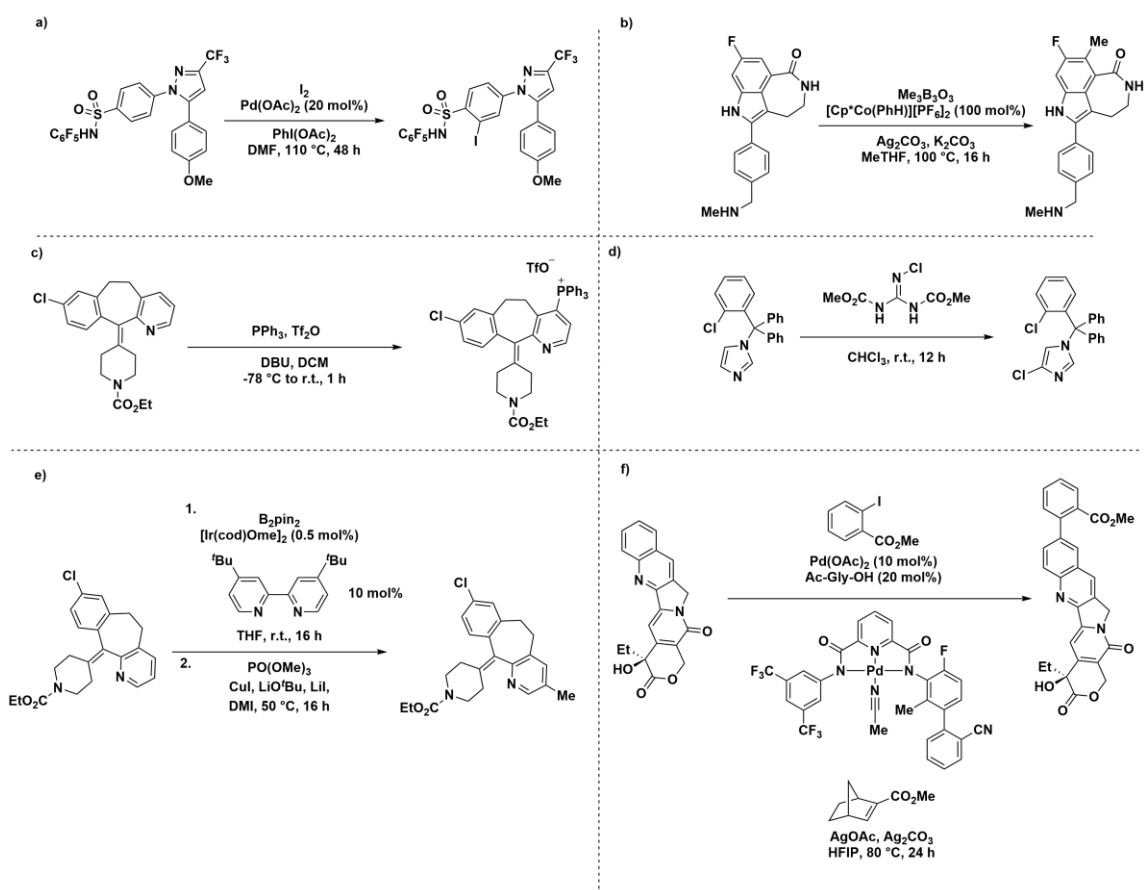
In the same vein as molecular property prediction, the site-specific reactivity of a particular compound can be predicted using machine learning.<sup>108</sup> This area of study is contributing to the field of retrosynthesis and synthetic route planning through the understanding of reaction behaviour without the need for experimental techniques.<sup>108–110</sup> This advancement in synthetic organic chemistry allows the rapid deployment of newly developed reactions to industry where the chemical space being studied is much larger than the reaction scope investigations carried out in initial research.<sup>111–114</sup> The range of applications includes the prediction of enantioselectivity in asymmetric catalysis,<sup>115</sup> diastereoselectivity of Diels-Alder reactions<sup>116</sup> and the focus of this work, the prediction of regioselectivity for an array of different reactions.<sup>117–119</sup> This is an exciting area of research since the ability to integrate new chemistry into the pharmaceutical industry quickly and effectively has the potential to have a profound impact on the drug discovery pipeline.

## 1.4 C-H Functionalisation

Since its inception, C-H functionalisation represents a new era of organic synthesis with its departure from functional conventional group manipulation. Instead, this approach relies on the selective functionalisation of specific C-H bonds under mild reaction conditions, allowing for a strong tolerance for other functional groups that may be present in the compound.<sup>120–124</sup> As the field of C-H functionalisation has

advanced it has become more sophisticated and can now be deployed in more complex molecules that contain multiple functional groups such as potential drug candidates.<sup>125</sup> This new era of functionalisation (known as late-stage functionalisation) has the advantage of reducing the number of steps in synthesis over classical synthetic methods, improving synthetic efficiency.<sup>126</sup> This ease of diversification of drug-like scaffolds allows facile structure-activity relationship exploration, modulation of pharmacokinetic properties to aid bioavailability and/or metabolism and installation of new handles for further functionalisation without the need for modification of synthetic route.<sup>127–129</sup> There are many methods of C-H functionalisation, but selectivity is dependent on three key determinants of reactivity. The first is the presence of directing groups in the compound. The second is the innate reactivity of the substrate, which can be influenced by the steric and electronic features of the molecule. Lastly, selectivity can be under reagent or catalyst control, where the electronic/conformational features of these components of the reaction determine regioselectivity.<sup>130</sup> Some examples of directing group functionalisation include the use of a pentafluoroaniline-derived sulfonamide to iodinate the position *ortho* to this group despite the presence of the strongly coordinating pyrazole functional group (**Figure 1.3a**);<sup>131</sup> or the cobalt(III)-catalysed C-H methylation of natural products (**Figure 1.3b**).<sup>132</sup> Innate selectivity examples include the C4-selective synthesis of phosphonium salts by triflic anhydride-mediated pyridine activation (**Figure 1.3c**),<sup>133</sup> and the selective C4 chlorination of the imidazole core of clotrimazole (**Figure 1.3d**).<sup>134</sup> Catalytic control examples include the one-pot iridium- and copper-catalysed methylation of loratadine (**Figure 1.3e**)<sup>135</sup> and the

palladium(II)-catalysed functionalisation of quinolines and isoquinolines such as camptothecin (**Figure 1.3f**).<sup>136</sup> While there are many different categories of C-H functionalisation, the primary focus of this work is in the area of functionalisation of heterocyclic species due to its particular value to the agrochemical and pharmaceutical industries.<sup>137</sup>

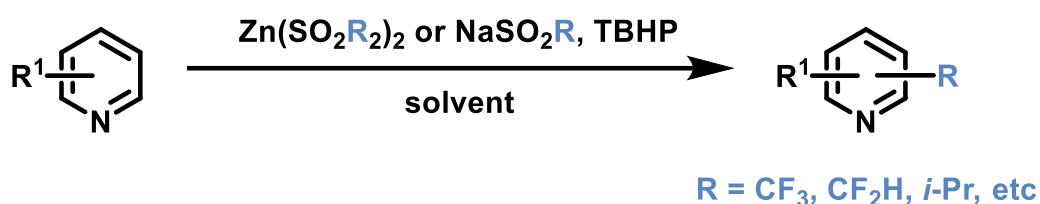


**Figure 1.3** Examples of C-H functionalisation with varying selectivity-determining processes.

The reaction of interest in our study is the functionalisation of aromatic heterocycles using a metal sulfonate reagent. Langlois *et al.*<sup>138</sup> originally developed sodium trifluoromethanesulfinate (Langlois reagent) as an effective trifluoromethylating agent for heteroarenes. This was revisited by Baran and co-workers,<sup>124,139–144</sup> who



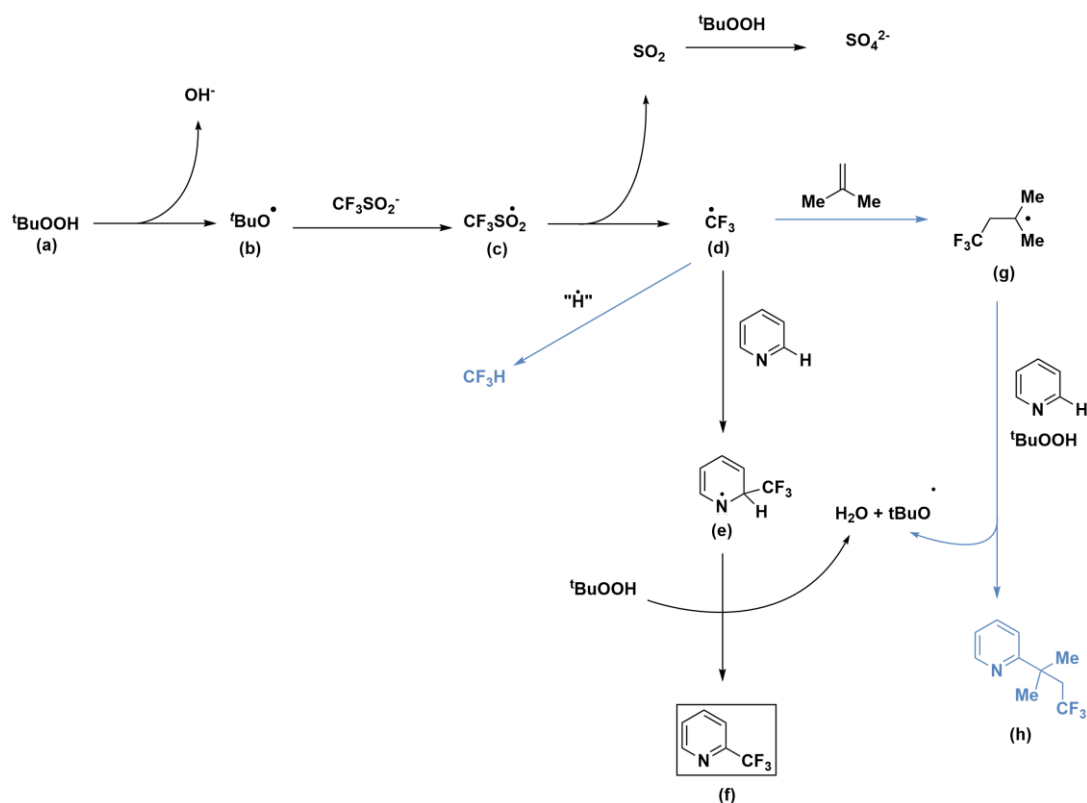
found that zinc sulfinate salts were similarly effective in C–H functionalisation but offered increased reactivity and bench-stability over sodium derivatives. This collection of zinc sulfinate salts is known as the Baran Diversinates. This method of functionalisation is attractive to organic chemists, due to its generalisability across a wide array of possible substitutions, provided that the required diversinate is available (**Figure 1.4**).<sup>145,146</sup> The reaction conditions reported by Baran *et al.* are also milder than those in Minisci<sup>147,148</sup> and Borono-Minisci<sup>149,150</sup> reactions, which require higher temperatures, strong oxidising agents and the use of expensive metal reagents such as AgNO<sub>3</sub>. This class of C–H functionalisation reaction is attractive to the field of drug discovery as it offers the ability to optimise a lead compound through substitution of the diversinate used in the reaction in a process known as late-stage functionalisation. This approach enables slight modifications to the structure of their compound that can be screened against their target of interest without the need for modification of the entire synthetic route to the parent compound.



**Figure 1.4.** Oxidative C–H functionalisation reactions of heteroarenes using zinc and sodium sulfinate salts.

A proposed mechanism (**Figure 1.5**)<sup>124</sup> involves the reduction of the *tert*-butylhydroperoxide oxidant (**a**) by trace metals to form the oxygen-centred radical <sup>t</sup>BuO• (**b**), which oxidises the trifluoromethanesulfinate anion (**c**). This generates the

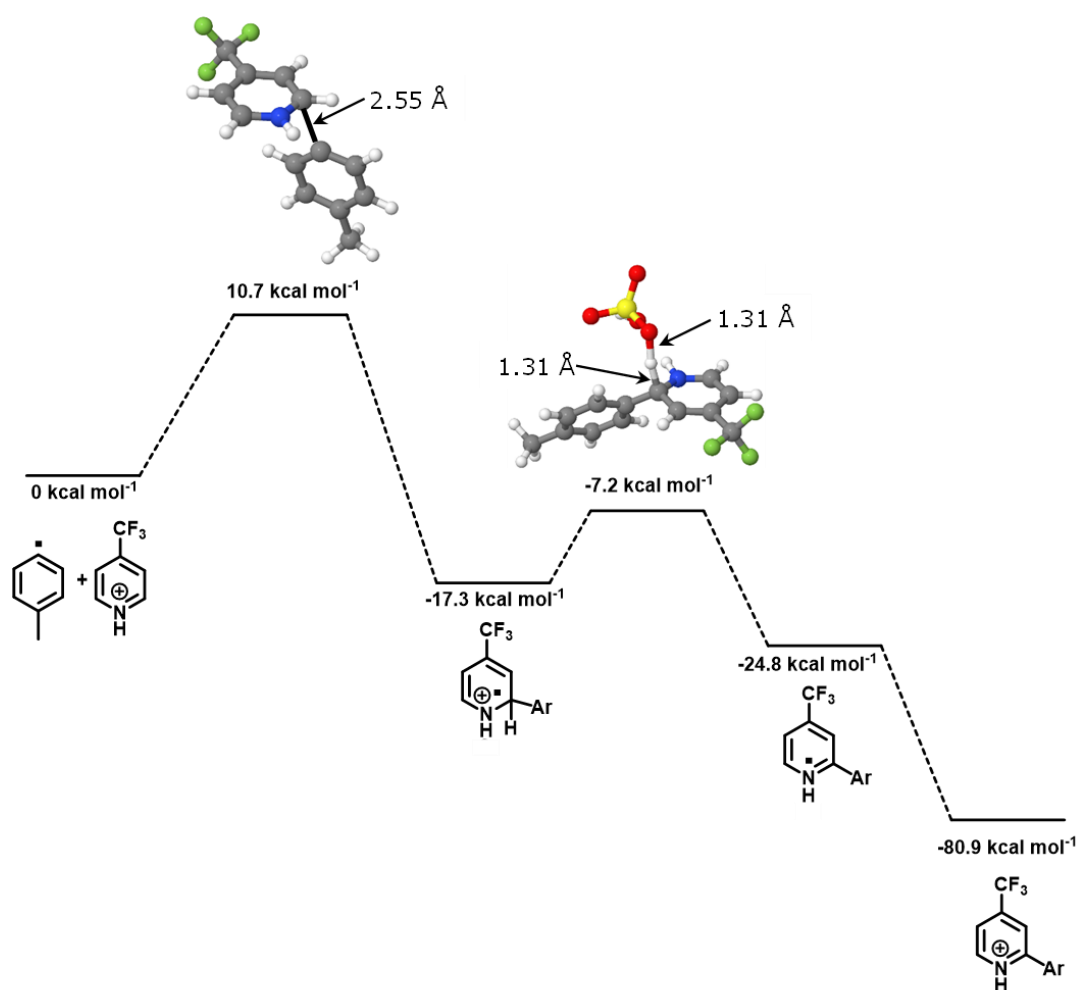
trifluoromethyl radical (**d**) upon release of SO<sub>2</sub>, which propagates through addition to the heteroarene (**e**) and *tert*-butyl peroxide-mediated re-aromatisation of the heterocycle to afford the desired product (**f**). An *isobutene* by-product (**g**) from sulfate anion formation can act as a radical trap, generating a new carbon radical which can then participate in Minisci-type addition to the aromatic moiety giving a second functionalised aromatic species as a side-product (**h**).<sup>151</sup>



**Figure 1.5:** Mechanism for metal sulfinate-mediated C-H functionalisation

A detailed kinetic, spectroscopic and density functional theory (DFT) study<sup>152</sup> of this functionalisation reaction showed that the attack of the radical on the heteroarene is the rate-limiting step (**Figure 1.6**) and that regioselectivity is kinetically controlled. However, the highly reactive nature of the attacking carbon radical intermediate makes the prediction of the position of radical addition challenging. While the radical

studied in this work differs from the arene species analysed in the study, the rate-limiting step is consistent between diversinates.



**Figure 1.6:** Free energy diagram for each step in zinc sulfinate-mediated C-H functionalisation computed at CPCM(water)-M06-2X/6-311+G(d,p)// B3LYP/6-31+G(d) level of theory. Figure reproduced from reference.<sup>152</sup>

## 1.5 Project Goals

This project aims to provide an effective method for predicting regioselectivity for this C-H functionalisation reaction through the application of quantum chemistry and machine learning. Through computational chemistry investigations, an effective method for predicting regioselectivity needs to be identified that balances accuracy with computational cost. Since there are only a select number of experimental examples with which to compare computational regiochemical predictions, calculation accuracy across these examples needs to be high to have confidence in the method's predictive ability on new areas of chemical space.

In order to utilise machine learning to effectively predict this regioselective behaviour, thousands of datapoints must be used to train the model to capture the relationship between descriptor and output. Therefore, the lack of experimental data motivates efforts towards the generation of an artificial dataset of a wide range of chemically different compounds. An accurate predictive model with a large domain of applicability is of great interest to drug discovery as this increases confidence in model output no matter the input compound. In order to model chemical processes/properties effectively, it is imperative that high-quality data is used. Publications in chemistry are heavily biased towards successful reactions, posing a problem for machine learning algorithms which require both positive and negative examples to understand the relationship between structure and experimental observation. Also, depending on the objective of the experiment, insufficient reporting of the full reaction conditions and side products leads to an incomplete

picture of the interactions between reagents in experiment and can therefore cause inaccurate predictions from the machine learning algorithm.<sup>153</sup> Artificial datasets have an advantage over data gathered from publications as they remove this experimental noise and isolate the key factors influencing regioselectivity.

Generation of this dataset requires large-scale calculation on an array of different compounds, leading to the development of an automated workflow to run and monitor these calculations in an efficient manner. Once this dataset has been generated it will then be used to build machine learning models to predict the regioselectivity, providing utility to the drug discovery industry.

In chapter 2 we discuss the theory behind two key areas of research used in this work, quantum chemistry and machine learning. In chapter 3 we examine the various quantum chemistry methods used to attempt to predict the regioselectivity of this C-H functionalisation reaction. In chapter 4 we discuss the computational workflow known as Rega which was developed for the generation of an artificial dataset for this reaction. In chapter 5 we discuss the deployment of Rega on a selected dataset and the generation of a variety of machine learning models to predict regioselectivity.

# Chapter 2 - Background Theory

There are two key fields of study throughout this work, quantum chemistry and machine learning. In order to understand the advancements in these areas covered in this research, it is important to familiarise oneself with the foundations upon which this work was built.

## 2.1 Quantum Chemistry

In this section, we discuss the fundamental principles of quantum chemistry, including the foundational Schrödinger wave equation and the approximations used to apply it to real-world chemical systems for the calculation of energies.

### 2.1.1 The Schrödinger Equation

The field of quantum chemistry is the study of quantum mechanics to explain relevant phenomena in chemistry. This may include the study of electronic configuration, bonding and potential energy surface exploration. “Modern” quantum mechanics was established in the 1920s, with the major breakthrough being the Schrödinger wave equation:<sup>154</sup>

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

where  $\Psi$  represents the wavefunction of the system, which describes the nature of the particles in the system,  $\hat{H}$  represents the molecular Hamiltonian operator and describes the energy of the associated electrons and nuclei in the system; and  $E$  is

the energy. The molecular Hamiltonian can be broken down further into kinetic and potential energy operators:

$$\hat{H} = \hat{T} + \hat{V} \quad (2.2)$$

where the kinetic energy operator  $\hat{T}$  describes each particle's kinetic energy term derived from its momentum:

$$\hat{T} = -\frac{\hbar^2}{2m} \frac{d^2}{dr^2} \quad (2.3)$$

where  $m$  represents the particle's mass. The potential energy operator  $\hat{V}$  describes the interactions of electrons and nuclei in the system using Coulomb's law:

$$E = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r} \quad (2.4)$$

where  $q_1 q_2$  represents the charge of the two interacting particles and  $r$  represents the distance between them. These equations can be simplified using atomic units, where  $m_e$  (electron mass),  $e$  (charge of an electron),  $a_0$  (Bohr's radius, nuclei-electron distance),  $\hbar$  (reduced Planck constant) and  $4\pi\epsilon_0$  (vacuum permittivity) are set to 1, yielding a simplified Hamiltonian accounting for all kinetic energy and Coulombic attraction/repulsion terms:

$$\begin{aligned} \hat{H} = & -\sum_I \frac{1}{2M_I} \nabla_I^2 + \sum_I \sum_{J>I} \frac{Z_I Z_J}{|R_I - R_J|} - \sum_i \frac{1}{2} \nabla_i^2 - \sum_i \sum_I \frac{Z_I}{|r_i - R_I|} \\ & + \sum_i \sum_{j>i} \frac{1}{|r_i - r_j|} \end{aligned} \quad (2.5)$$

where the first term describes the sum of the nuclear kinetic energy of the nuclei, the second term describes the sum of the nuclear-nuclear repulsion, the third term is the sum of the electronic kinetic energy, the fourth term is the sum of the electron-nuclear attraction, and the final term is the sum of the electron-electron repulsion.

As can be seen above, the Schrödinger equation becomes very complex when all factors influencing the energy are accounted for. In fact, for a many-body system the Schrödinger equation becomes impossible to solve exactly. Therefore, approximations are employed to simplify the problem. Firstly, the Born-Oppenheimer approximation<sup>155</sup> assumes that the electronic structure of the molecule adjusts instantaneously to shifts in the position of the nuclei, giving two separate Schrödinger equations for nuclear and electronic motion. Since in quantum chemistry, the primary concern is the electronic structure, the Hamiltonian is simplified through removal of the nuclear kinetic energy term, leaving only electronic terms in the operator known as the electronic Hamiltonian. This is simplified further by molecular orbital theory, which approximates the exact wavefunction of the system by constructing molecular orbitals. The variational principle dictates that the approximate wavefunction of the system will be greater in energy than the exact wavefunction, giving rise to an approximate energy of the system. The molecular orbitals comprising the approximate wavefunction can be tuned to give as low an energy as possible to get as close to the exact wavefunction as possible. These molecular orbitals are comprised of a linear combination of atomic orbitals and take the form:



$$\psi = \sum_i a_i \phi_i \quad (2.6)$$

where  $\psi$  represents a single molecular orbital,  $a_i$  is the molecular orbital coefficient and  $\phi_i$  is the atomic orbital. Alterations in size and complexity of the atomic orbitals through the use of different basis sets give rise to different approximations to the exact electronic wavefunction of the system. Basis sets describe the size and shape of the atomic orbitals and includes different approximations to account for potential interactions between electrons in a molecule. The basis set chosen is based on the computational cost and accuracy required to effectively simulate the system in question. One family of basis sets used in this work is the split-valence set.<sup>156</sup> This family use a greater number of basis functions to describe the valence orbitals than the core orbitals since these electrons are used for bonding and take part in the modelling of chemical reactions. The 6-31G\* split valence basis set uses one basis function comprised of six Gaussian functions to describe the core electronic orbitals; and two basis functions comprising three Gaussians and one Gaussian respectively to describe the valence electronic orbitals. While the use of Gaussian functions (2.7) in this family is not as accurate a depiction of orbital shape/size as the Slater-type Orbital (STO)<sup>157</sup>(2.8), the exponential term in the function allows for rapid evaluation of the two-electron integrals:

$$g(r) = e^{-\zeta_v r^2} = e^{-\zeta_v(x^2+y^2+z^2)} \quad (2.7)$$

$$S_v(r) = e^{-\zeta_v |r|} = e^{-\zeta_v \sqrt{x^2+y^2+z^2}} \quad (2.8)$$

where zeta ( $\zeta$ ) controls the shape of the function and  $(x^2 + y^2 + z^2)$  represent Cartesian space. The combination of multiple gaussian functions allows for effective

mapping of the STO while improving calculation efficiency. Another basis set family used in this work is the Karlsruhe set.<sup>158–161</sup> This family of basis sets use split valence approach comprising of gaussian basis functions with varying levels of splitting depending on the desired accuracy. Additional functions can be added that describe the atomic orbitals' polarizability and diffusion. In this work, def2-TZVP is used, which uses a triple zeta ( $\zeta$ ) valence orbital with added polarisation function to accurately model the atomic orbital shape. When calculating the energy of the molecule using these basis sets, a fixed set of atomic orbitals allows optimisation of the molecular orbital coefficients  $a_i$  and therefore the minimisation of the energy through the application of the variational principle.

Assembly of the molecular orbitals through modification of the molecular orbital coefficients is challenging because the energy of the system is dependent on the wavefunction of the system. To circumvent this, an iterative process known as the self-consistent field (SCF) method is applied. Initially, a “guess” set of molecular orbitals is chosen to represent the molecule which are then used to find the energy of the system. Then a new set of molecular orbitals is constructed, and the energy of that system is compared with the first. This process continues until the energy of the system is minimised and the best set of molecular orbitals to represent the wavefunction of the system has been found.

## 2.1.2 Hartree-Fock Theory

The simplest *ab initio* method (clarification into approximations used given in 2.1.4) is Hartree-Fock theory,<sup>162,163</sup> which utilises this variational principle to determine the orbital energies comprising the approximate electronic wavefunction. It does this through the understanding of electron spin and the interactions of these electrons to construct a spin orbital:

$$\chi^\uparrow(x) = \psi(r)\alpha(\omega) \quad (2.9)$$

$$\chi^\downarrow(x) = \psi(r)\beta(\omega) \quad (2.10)$$

where  $\chi$  are the spin orbitals,  $\psi(r)$  is the spatial orbital and  $\alpha(\omega)$  and  $\beta(\omega)$  are the spin functions for spin up and spin down orbitals respectively. This gives rise to the construction of the N-electron wavefunction of the system, comprised of these spin orbitals through the use of a Slater determinant:

$$\Psi_{SD}(x_1 \dots x_N) = \frac{1}{\sqrt{N!}} \begin{bmatrix} \chi_1(x_1) & \cdots & \chi_N(x_1) \\ \vdots & \ddots & \vdots \\ \chi_1(x_N) & \cdots & \chi_N(x_N) \end{bmatrix} \quad (2.11)$$

where  $x_1 \dots x_N$  are the space, spin coordinates for each electron in the N-electron system. The electronic Hamiltonian acting on this Slater determinant gives the Hartree-Fock energy. The factors influencing each electron are the kinetic and nuclear attraction energy, the Coulomb interaction (energy of each electron repelling one another) and the exchange interaction energy (a purely quantum mechanical effect of electrons of the same spin in the same molecular orbital interacting with one another). Application of the variational principle in the Hartree-Fock method gives the following:

$$\hat{f}\psi_i = \varepsilon_i\psi_i(r) \quad (2.12)$$

where  $\hat{f}$  is the Fock operator (which includes the aforementioned electronic interaction terms),  $\psi_i$  is the molecular orbital and  $\varepsilon_i$  is the orbital energy. This gives the energy and molecular orbital for each electron in the system. Of the interaction terms included in the Fock operator, the two-electron Coulomb and Exchange integrals are the most time computationally expensive, with the time taken scaling with the number of basis functions included in the atomic orbitals to the power of 4:

$$T_{HF} = Nb^4 \quad (2.13)$$

due to the evaluation of the four-centre two-electron integrals. In these integrals, each electron is described by a linear combination of basis functions centred over multiple atoms, giving  $Nb^4$  scaling.

### 2.1.3 Semi-empirical Methods

When performing calculations on very large systems such as proteins, the cost of *ab-initio* methods such as Hartree-Fock becomes impractical. Therefore, further simplification of the most costly elements of these calculations is required to make simulation of large systems tractable. In semi-empirical computational methods, explicit calculation of the two-electron integrals is avoided and replaced with tabulated values based on either experimental or previously calculated computational data. There are a number of different semi-empirical methods, each with different approximations and parameterisations. One of the most common methods is the Austin-Model 1 (AM1)<sup>164</sup> which utilises the neglect of differential

diatomic overlap (NDDO) approximation<sup>165</sup> to avoid the calculation of the electron-electron repulsion terms and parameterisation of integrals based on experimental measurement of dipole moments and ionisation potentials. Another popular method is the parametric method 7 (PM7)<sup>166</sup> which is parameterised based on high-level *ab-initio* calculation in order to reduce the error compared to these more expensive methods, increasing confidence in the output. In later chapters both semi-empirical methods were screened against predicting regioselectivity in our reaction of interest. AM1 had greater predictive performance than PM7 when compared to experiment and so it was chosen for subsequent calculations in the automation workflow described in later chapters.

## 2.1.4 Density Functional Theory

While Hartree-Fock is shown to be more accurate than other simpler approximations such as semi-empirical methods, it does not completely account for all the interactions that can occur between two electrons. Thus, electrons are typically placed too close together, raising the overall energy of the system compared to the ground truth. This error in Hartree-Fock is known as the correlation energy:

$$E_C = E_{exact} - E_{HF} \quad (2.14)$$

where  $E_C$  is the correlation energy,  $E_{exact}$  is the exact energy for a given basis set and  $E_{HF}$  is the Hartree-Fock Energy. The correlation energy is due to two distinct electronic effects, the dynamical correlation which arises from the poor description of short-range instantaneous interactions between electrons, and static correlation,

which describes the long-range interactions between electrons not fully described by the Slater determinant used in Hartree-Fock. Some more expensive methods such as coupled cluster theory account for this correlation energy through perturbation of the Hartree-Fock wavefunction, generating determinants for different excitation levels for each electron and factoring the effects of these excited electron interactions into the energy. A common coupled cluster theory utilised is CCSD,<sup>167</sup> which accounts for singly and doubly excited electron interactions. This method is very accurate but is much more costly:

$$T_{CCSD} = Nb^6 \quad (2.15)$$

due to the iterative process required to initially calculate the Hartree-Fock two-electron integrals plus the energies of electrons in the excited Slater determinant. Another method used to include these correlation effects is Møller-Plesset 2 (MP2).<sup>168</sup> In this method, a perturbation is applied to the Hartree-Fock Hamiltonian to model the shifts in energy levels of the electrons caused by correlation effects. In MP2 theory, the 2 signifies a second order correction to the Hartree-Fock energy to include these correlation effects. However, this method can still be deemed too costly for many-electron systems:

$$T_{MP2} = Nb^5 \quad (2.16)$$

where the second order correction adds another layer of expense to the  $Nb^4$  scaling of Hartree-Fock. To circumvent this problem, an alternative approach was developed that accounted for correlation energy at a similar base computational cost to Hartree-Fock. Density Functional theory<sup>169</sup> proposed that instead of studying the

many-electron wavefunction which depends on  $3N$  coordinates, the electron density of the system is used which only relies on 3 coordinates regardless of the system size:

$$\rho(r_1) = N \int |\Psi(x_1, x_2 \dots x_N)|^2 d\omega_1 dx_2 \dots dx_N \quad (2.17)$$

where  $\rho(r_1)$  is the electron density of electron 1,  $N$  is the number of electrons in the system,  $|\Psi(x_1, x_2 \dots x_N)|^2$  is the probability density for the wavefunction over all coordinates except for  $d\omega_1 dx_2 \dots dx_N$ , the spatial coordinates for electron 1. This expression describes the probability of finding any electrons in the volume element  $dr_1$ .

In density functional theory the nuclear attraction term on the electronic Hamiltonian is referred to as the external potential. Two theorems established by Hohenberg and Kohn provided a foundation for the field of electron density study. The first theorem is that the electron density determines the external potential of the system<sup>170</sup>. Therefore, from the electron density, we can determine the full Hamiltonian, the wavefunction of the system (via the Schrödinger equation) and all ground state properties of the system. The ground state energy can be written as a functional of the density:

$$E[\rho] = E_{NE}[\rho] + T[\rho] + W[\rho] = \int \rho(r)V(r)dr + F[\rho] \quad (2.18)$$

$$F[\rho] = T[\rho] + W[\rho] \quad (2.19)$$

where  $E_{NE}[\rho]$  is the nuclear-electron attraction given by the density,  $T[\rho]$  is the electron kinetic energy term and  $W[\rho]$  is the electron-electron repulsion term.  $F[\rho]$  is the universal density functional and includes the electron kinetic energy and

repulsion terms. The second theorem is that any approximate electron density gives an energy greater than or equal to the actual energy. While the density could be obtained from the physical system where all electrons interact with one another, this approach would require the many-electron wavefunction which is trying to be avoided. Instead, Kohn and Sham proposed generating an initial electron density from a system of non-interacting particles, which can be obtained from a single Slater determinant similar to Hartree-Fock. In doing so we could obtain the exact energy of the system using electron density:

$$E[\rho] = E_{NE}[\rho] + T_s[\rho] + J[\rho] + (T[\rho] - T_s[\rho] + W[\rho] - J[\rho]) \quad (2.20)$$

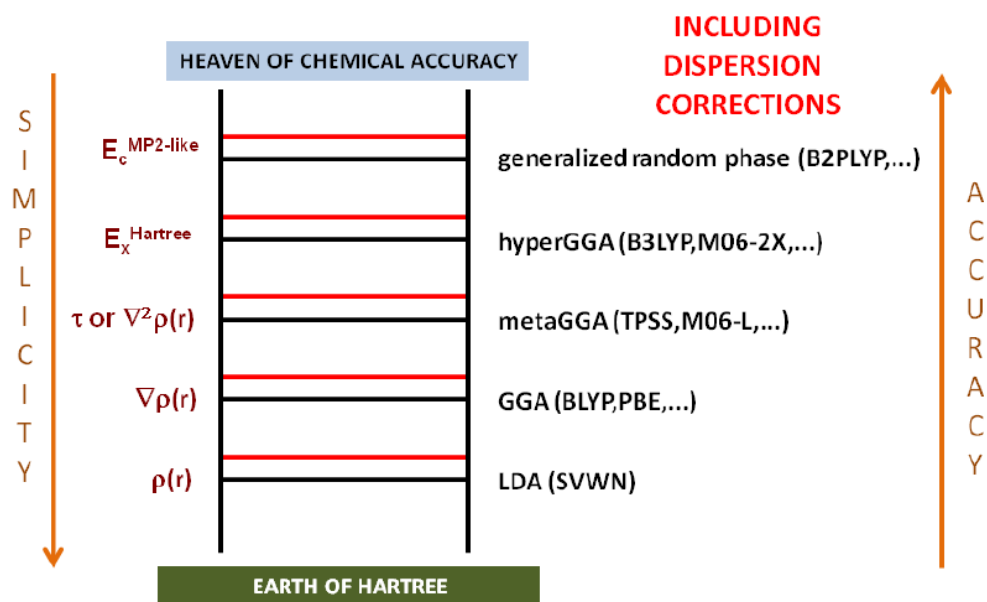
where  $T_s[\rho]$  is the kinetic energy functional for the non-interacting system,  $J[\rho]$  is the Coulomb energy,  $T[\rho] - T_s[\rho]$  are the terms describing the difference in kinetic energy between the interacting and non-interacting system and  $W[\rho]$  is the electron-electron repulsion functional. The final four terms are collectively known as the exchange-correlation functional and account for the difference between the single Slater determinant derived non-interacting electron terms and the true interacting system:

$$E[\rho] = E_{NE}[\rho] + T_s[\rho] + J[\rho] + E_{XC}[\rho] \quad (2.21)$$

Since this exchange-correlation functional cannot be known exactly, it is approximated via a variety of different functionals with varying levels of complexity. This is known as the Jacob's ladder of functionals (**Figure 2.1**).<sup>171</sup> The simplest and lowest "rung on the ladder" is the local density approximation (LDA) class of functionals,<sup>170</sup> which assumes the electron density across the molecule is uniform



and so the counterbalancing positive charge of the nuclei is spread evenly across the system. This was not an accurate representation of the charge distribution seen in a molecule and so the next “rung on the ladder” attempted to solve this. The first functional showing promising applications in describing experiment was the generalised gradient approximation (GGA),<sup>172</sup> which applies the positive charge of the system to the electron density based on the gradient of electron density to account for the ununiform electron density present in a molecule. The next most sophisticated class of functionals is known as the meta-GGA (mGGA) functionals which attempt to improve upon the GGA class by including more non-local information in the charge distribution through the use of the second derivative of the electron density known as the Laplacian.<sup>173</sup> This was built upon substantially on the final “rung” to be discussed with the development of hybrid functionals which added a fraction of the Hartree-Fock orbital dependent exchange to the GGA energy.<sup>174</sup> Today, the most commonly used functionals in chemistry are hybrid functionals, namely B3LYP<sup>175,176</sup> and the Minnesota class of functionals such as M06-2x.<sup>177</sup>



**Figure 2.1** Jacob's ladder approach for the systematic improvement of DFT functionals. Figure reproduced from reference.<sup>178</sup>

### 2.1.5 Atom-centred charges and Fukui Indices

One calculated property of interest to the task of regioselectivity prediction is the calculation of atom-centred charges. In this area of study, the charge of each nuclei is partitioned based on the surrounding electron density, and a compound's regioselectivity can be measured from this distribution of atomic charge. A commonly used example of atom-centred charge calculations is Mulliken charges,<sup>179</sup> which calculates the partial atomic charge of an atom and its contribution to the overall charge of the molecule, based upon the linear combination of atomic orbitals. One of the charge methods used in this work is Hirshfeld charges,<sup>180</sup> which defines the charge contribution  $q$  of atom  $X$  as:

$$q_X = Z_X - \int \frac{\rho_X^0(r)}{\sum_Y \rho_Y^0(r)} \rho(r) dr \quad (2.22)$$

where  $Z_X$  is the atomic number of the element  $X$ ,  $\rho$  is the molecular density and  $\rho_X^0$  is density of  $X$  as an isolated atom. This approach to the partitioning of atomic charge is less dependent on the choice of basis set compared with Mulliken charges. Another charge method used in this work is the atomic dipole-corrected Hirshfeld charge (ADCH),<sup>181</sup> which account for the deficiencies of Hirshfeld charges by including the atomic dipole moment of each atom in the calculation.

Another calculated property of interest to this work are Fukui indices. These Fukui indices are based on the Fukui function<sup>182</sup> which describes the change in electron density at a given position when the number of electrons has been altered:

$$f(r) = \frac{\partial \rho(r)}{\partial N} \quad (2.23)$$

where  $\rho$  is the electron density and  $N$  is the number of electrons in the system. This corresponds to the change in reactivity of the molecule at a point in space due to alteration of the frontier orbitals, namely the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). Since the focus of predicting site selectivity is investigating the difference in reactivity on an atom-by-atom basis, the condensed Fukui function describes the change in electron density around each atom and their contributions to the change in overall electron density. The charge scheme used for this calculation is known as the finite difference method, where the electron density of the system is calculated upon addition and removal of an electron and the change in each atomic charge to balance the system is given by the following:

$$f_{v,N}^{+, \alpha} = \rho_{v,N+1}^{(\alpha)} - \rho_{v,N}^{(\alpha)} \quad (2.24)$$

$$f_{v,N}^{-, \alpha} = \rho_{v,N}^{(\alpha)} - \rho_{v,N-1}^{(\alpha)} \quad (2.25)$$

$$f_{v,N}^{0, \alpha} = \frac{1}{2} \left( \rho_{v,N+1}^{(\alpha)} - \rho_{v,N-1}^{(\alpha)} \right) \quad (2.26)$$

where  $\rho^{\alpha}$  describes the electron density around atom  $\alpha$  and  $f_{v,N}^{+, \alpha}$ ,  $f_{v,N}^{-, \alpha}$ , and  $f_{v,N}^{0, \alpha}$  correspond to the atom's susceptibility to nucleophilic, electrophilic and radical attack respectively. This atom-centred property could prove useful in the prediction of regioselectivity. While studies have shown varying success in the prediction of site-specific reactivity depending on the division of electron density throughout the molecule,<sup>183</sup> its' efficacy will be considered in this investigation.

## 2.1.6 Transition State and Frequency Calculation Procedure

The above methods describe the calculation of the energy of the system based on the construction of the molecule and consideration of the interaction terms that contribute to the overall energy. In order to locate either the optimised geometry or a transition state, the energy is computed for a given conformation and the force is calculated:

$$F = - \frac{dE}{dR} \quad (2.27)$$

where  $R$  is the interatomic distance for each atom in the molecule. This force describes the nature of the potential energy surface (PES), the multi-dimensional relationship between bond lengths and angles and the corresponding energy. The number of dimensions of the potential energy surface is  $3N - 6$  where  $N$  is the

number of nuclei. This corresponds to all of the vibrational modes the molecule possesses. The force gives the gradient of the slope at a particular point on the PES and informs the QM software package of the alterations in geometry required to locate a stationary point on the surface, where  $F = 0$ . This procedure is followed for the location of both minimum energy conformations as well as transition states. When determining the nature of the stationary point, the vibrational frequencies  $\nu_k$  and vibrational modes  $\eta_k$  are given by the eigenvalue equation:

$$\tilde{H}\eta_k = \lambda_k\eta_k \quad (2.28)$$

where  $\tilde{H}$  is the mass-weighted Hessian:

$$\tilde{H}_{IJ} = \frac{1}{\sqrt{m_I m_J}} \nabla^2 E(\mathbf{R}^*) = \begin{bmatrix} \frac{1}{\sqrt{m_1 m_1}} \frac{\partial^2 E}{\partial R_1 \partial R_1} & \cdots & \frac{1}{\sqrt{m_1 m_n}} \frac{\partial^2 E}{\partial R_1 \partial R_n} \\ \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{m_n m_1}} \frac{\partial^2 E}{\partial R_n \partial R_1} & \cdots & \frac{1}{\sqrt{m_n m_n}} \frac{\partial^2 E}{\partial R_n \partial R_n} \end{bmatrix} \quad (2.29)$$

where  $n$  corresponds to each of the nuclear coordinates. The eigenvalues of the molecular Hessian are then directly related to the vibrational frequencies through the below:

$$\nu_k = \frac{\lambda_k^{\frac{1}{2}}}{2\pi} \quad (2.30)$$

where  $\lambda_k$  represents a given eigenvalue in the matrix. In order to determine the nature of the stationary point the inertia of the matrix can be used:

$$\text{In}(\mathbf{M}) = [\pi(\mathbf{M}), \nu(\mathbf{M}), \delta(\mathbf{M})] \quad (2.31)$$

where matrix  $\mathbf{M}$  is equal to a list of the number of positive eigenvalues  $\pi$ , the number of negative eigenvalues  $\nu$  and the number of zero eigenvalues  $\delta$ . For  $n$  vibrational modes, a minimum on the PES has an inertia of  $[n, 0, 0]$ , meaning zero negative and zero eigenvalues. For a maximum on the PES, the inertia is  $[n - k, k, 0]$  where  $k$  is the  $k^{th}$  order saddle point. For a transition state where the surface is negative in one direction (the direction corresponding to the reaction coordinate) and positive in all other directions and so  $k = 1$ . The single negative eigenvalue gives an imaginary frequency and so the presence of this frequency in the calculation output can be used to determine whether the optimisation has converged to a TS.

### 2.1.7 Solvent Models and Other QM Calculation Techniques

When modelling experiment, the choice of solvent may have a large impact on the favoured reaction pathway. Since these solvents interact with the substrate, they may redistribute electron density across the molecule leading to a difference in the prediction of the most favourable product when compared to the calculation of the substrate in a vacuum. This can be accounted for in QM calculation through the addition of solvation models. There are two types of solvation models, implicit and explicit. In explicit solvation models, solvent molecules are directly added into the system being calculated and are allowed to interact with the substrate. These models are more accurate than implicit methods but come at the cost of increased computational expense and as such are typically reserved for less intensive QM methods such as molecular dynamics and molecular mechanics. Implicit solvent models aim to approximate the interactions between solvent and the molecule of

interest by applying a correction to the energy by adding a continuously polarisable medium of a given dielectric constant to represent the interactions between solvent and solute.<sup>184</sup> One such example of an implicit solvation model is the solvation model based on density (SMD) method. In this model, the interaction between the dielectric constant of the solvent and the full electron density of the solute is measured to calculate the solvation energy of the system.<sup>185</sup>

There are other processes used in QM methods that are used to refine the energy calculation procedure. When performing a geometry optimisation or transition state search, it is common practice to use a smaller basis set to perform the optimisation procedure followed by the calculation of the energy of the system using a larger basis set. This method reduces the computational expense compared with full optimisation with the larger basis set (considering the large number of SCF cycles typically required to perform the optimisation) and it is assumed that the difference in optimised conformation between basis sets is negligible. The subsequent larger basis set energy calculation then more accurately represents the total energy of the system. If this process is used, it is denoted by a “//” between the methods used, with the optimisation level of theory preceding the dashes and the energy calculation level of theory appearing after.

## 2.2 Machine Learning

In this section, we discuss the various machine learning techniques employed in this work and some of the descriptor sets used in the generation of these models.

### 2.2.1 Classical QSAR Modelling Methods

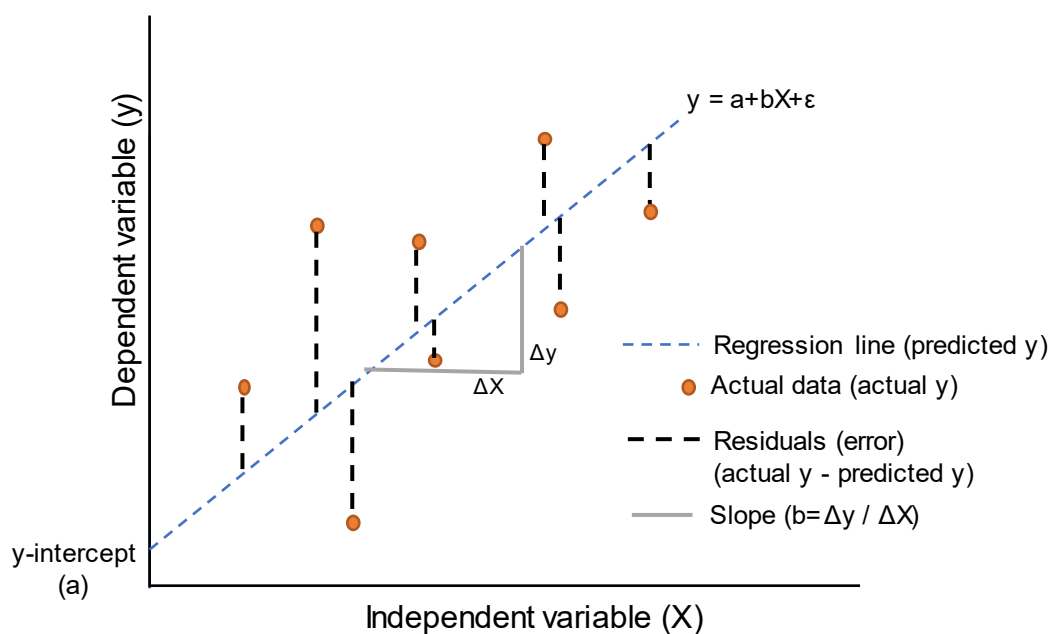
The first applications of machine learning in chemistry attempted to predict molecular properties through the relationship between the compound's structure and its activity, known as quantitative structure-activity relationship (QSAR).<sup>186</sup> This approach is prevalent in drug discovery since in this area a drug is typically designed around its conformation and therefore its binding ability to the target receptor. Therefore, information garnered about the molecule's shape, size and varying local environments provide insight into its interaction with the receptor. QSAR models attempt to either predict a quantitative property (such as binding affinity) in a regression-based task or categorial value (such as site-specific metabolism) in a classification-based model. Model training utilises datasets split into training, validation and test sets. In the training set, the predicted variable is shown to the model so that the relationship between it and the molecular properties can be learned. In the validation set, the parameters of the model are tuned to predict the output variable on a different set of compounds, improving generalisability. In the test set, the real output variable value is withheld, and the model must give a prediction on the compounds in the set. The prediction accuracy is then measured against the real output value to assess the model performance. There are a variety



of different methods used to identify the relationship between structural properties and activity, some of which are outlined below.

## Linear Regression

Regression modelling estimates the linear relationship between the dependent variable and independent variable by finding a suitable fit between the data points (Figure 2.2)<sup>187</sup>.



**Figure 2.2** The relationship between the dependent variable (y) and the independent variable (x). The deviations of the data points (red) to the fit are due to other factors influencing the dependent variable not captured in x. Figure reproduced from reference.<sup>188</sup>

If  $y$  is the experimentally observed dependent variable and  $x$  is the descriptor, then the equation of the straight line relating these two variables is:

$$y = b + ax + \epsilon \quad (2.32)$$

where  $b$  is the intercept,  $a$  is the slope of the regressor gradient and  $\epsilon$  is the error term known as the residuals. The slope and intercept are evaluated to minimise the error term which is the difference between the regressor prediction and the actual datapoint. The values of  $a$ ,  $b$  and  $\epsilon$  are given by the following:

$$a = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (2.33)$$

$$b = \bar{y} - a\bar{x} \quad (2.34)$$

$$\epsilon = y - \hat{y} \quad (2.35)$$

where  $N$  is the number of data points,  $\bar{x}$  and  $\bar{y}$  represent the means of the independent and dependent variables and  $\hat{y}$  is the predicted value of  $y$ . This modelling technique is known as simple linear regression and only accepts one descriptor as a possible dependent variable. In the case of QSAR modelling, multiple linear regression is employed by including the many different descriptor categories found in chemistry to estimate the weighted contributions of these independent variables to the dependent variable:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i \quad (2.36)$$

where  $\beta_0$  represents the intercept of the regressor,  $\epsilon_i$  describes the disturbance term, a variable which captures all factors influencing the dependent variable  $Y_i$  other than  $X_{ip}$ ;  $\beta_1 \dots \beta_p$  describe the regression coefficients for  $p$  datapoints where  $i = 1 \dots p$ . Therefore,  $Y_i$  is the contribution of data points  $i$  to the regressor's fit. Linear regression modelling is one of the simplest forms of machine learning techniques and is typically used as a benchmark for other methods to improve since

other methods can gain a greater understanding of the commonly non-linear relationships seen between descriptor and output.

## **Random Forests**

Random forests modelling<sup>189–191</sup> is an ensemble technique that constructs a number of decision trees and aggregates the results of these trees in either a regression or classification task. Ensemble techniques are a powerful tool in machine learning where a number of “weak” models are trained on the same predictive task. These weak models may each have poor predictive ability through underfitting to the training data (known as bias) or a poor generalisability to unseen data through overfitting (known as variance). The aggregation and blending of each of these models’ outputs aims to properly fit the training data to improve predictive accuracy and improve model generalisability by balancing bias and variance. This can be done through a technique known as bagging or bootstrap aggregating, whereby the random selection of features used to train each weak model can identify the subtle relationship between some features and the output. This singular model in isolation would give poor model performance, however, the aggregation of the patterns these models identify can be used to learn the full relationship between all input features and the output value. In Random forests modelling the weak models used are decision trees. Decision trees<sup>192,193</sup> consist of a series of junctions where a descriptor value is used to learn the relationship between it and the output variable. For example, if an input descriptor is the compound’s Log P, the junction or node in the tree may consist of a specific Log P value. Therefore, when training the tree, the

compound's properties are probed at each junction and then related to the output variable. An example decision at a particular junction may be "all compounds that are active have a Log P value of greater than 5". Different descriptors are employed at different decision points throughout the tree and the model is considered trained when the values used at each decision point accurately predict the output variable within a certain threshold. Random forests generate a number of decision trees with a subset of randomly chosen descriptors used at each decision point and combine the predictive output to assess the majority vote. Doing so removes the possibility of overtraining, a problem whereby the model training set performance improves at the expense of generalisability to the test set.<sup>194–196</sup> In regression tasks, the mean predicted value of the decision trees is given as the random forest output. In classification tasks, the class selected by the greatest number of trees is the random forests predicted class. This method is still commonly used today, with its performance across different datasets and tasks being benchmarked against other machine learning methods.

## **XGBoost**

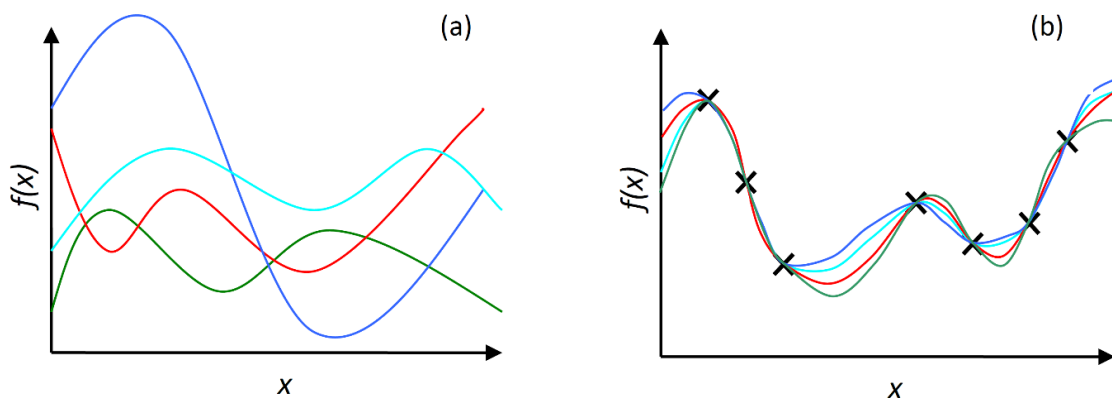
Gradient boosting is a machine learning technique which utilises "weak" learners such as decision trees and an additive model which combines the results of the weak learner models to minimise a loss function and inform the shape and size of the next weak learner in an iterative process. This approach is known as functional gradient descent and acts to maximise single model performance rather than employ an ensemble approach of randomly generated "weak" models such as random forests.

In the case of XGBoost,<sup>197</sup> decision trees are used as the “weak” learner and the first step is an initial prediction of tree shape/depth. At the end of the tree (known as the leaves), the similarity between compounds in the same leaf is calculated. The second tree builds upon this result by altering the descriptors and threshold values at each node of the tree. This new tree then has similarity calculated at each leaf and the gain in similarity from the previous tree is assessed. This process is repeated until the algorithm converges on an optimal decision tree structure where no additional gains in performance can be obtained. XGBoost has techniques built within the algorithm to avoid overfitting to the training data, such as pruning a “branch” of the tree where the gain in similarity between compounds is too small to reach a certain threshold (known as tree complexity parameter or gamma).

XGBoost is often regarded as an improvement on random forests modelling since it takes a non-random and repeatable approach to generating the best tree structure and node features. When making predictions, this method uses weighted contributions of each tree's predicted value based on its performance, rather than the average predicted value given by random forests.

## **Gaussian Processes**

Gaussian processes (GP) use a Bayesian probabilistic approach to inference whereby a prior probability distribution of an unknown function is assumed and then this distribution is updated (**Figure 2.3**) to generate a posterior distribution for functions that fit the observed data for the training points.



**Figure 2.3:** Random samples from (a) prior distribution of functions and (b) posterior distribution of functions in a one-dimensional example. Functions from the posterior distribution are conditioned to pass through the training points shown by crosses.

Gaussian process modelling places a distribution (known as a prior) over the function that fits the training points. If the descriptors are seen as a set of vectors, then the prior for the function that fits them is a multidimensional distribution with zero mean and covariance matrix  $Q$  that depends on these descriptor vectors (i.e. each descriptor pair has a covariance). The components of this matrix  $Q$  are given by the Gaussian covariance function  $C(\underline{x}_n \underline{x}_m)$ , which is used to define the distance between each element in the input (i.e. the similarity between molecules). An example of a covariance function is given below:

$$C(\underline{x}_n \underline{x}_m) = \theta_1 \exp \left[ -\frac{1}{2} \sum_{i=1}^K (x_{ni} - x_{mi})^2 / r_i^2 \right] + \theta_2 \quad (2.37)$$

where  $\underline{x}_n$  and  $\underline{x}_m$  are different points in the training data,  $\theta_1, \theta_2, r_i$  and  $i = 1 \dots K$  are hyperparameters.  $r_i$  is a set of length scale parameters; one for each

descriptor. If  $r_i$  is small, it represents a descriptor that has a strong influence on the observed property. The covariance function serves to determine the relationship between each descriptor and the output variable. Different covariance functions (also referred to as a kernel) may be applied to different data types to determine the shape of the prior and posterior distributions and hence can be altered to influence predictive performance. This method means that when trained, a new input compound  $y'$  with a new descriptor vector  $x'$  is presented to the model ( $y' = y(x')$ ), prediction confidence can be obtained in the form of a Gaussian distribution with mean (2.38) and variance (2.39):

$$\mu = k^T(Q + \theta_3 I)^{-1}Y \quad (2.38)$$

$$\sigma^2 = \kappa - k^T(Q + \theta_3 I)^{-1}k \quad (2.39)$$

where the vector  $k$  with components  $k_n = C(\underline{x}_n \underline{x}_m)$  describes the similarity of new molecules to that of the training set,  $\kappa = C(x', x')$  and  $\theta_3$  represent the variance in the assumed noise in the data and  $Y$  is the property output value. The mean of the Gaussian distribution is taken as the predicted property for the new molecule  $\mu$ . The standard deviation  $\sigma$  of this distribution is a measure of the uncertainty in the predicted value and can be used as an indicator of how different the new molecule  $\mu$  is in descriptor space from the training data. If  $\sigma$  is large it shows the new molecule is well outside the descriptor space of the training data and hence the uncertainty is greater.

Each of the hyperparameters  $\theta$  (2.40) are learned from the training data. To ensure the function is smooth, the most probable set of hyperparameters is chosen, by finding the maximum of the log marginal likelihood. This also ensures that the model is not over-trained as a compromise between curvature and level of fitting is achieved. Gaussian processes modelling is known for having a greater extrapolation ability compared to random forests methods due to its smoothing of the fitting between training points.

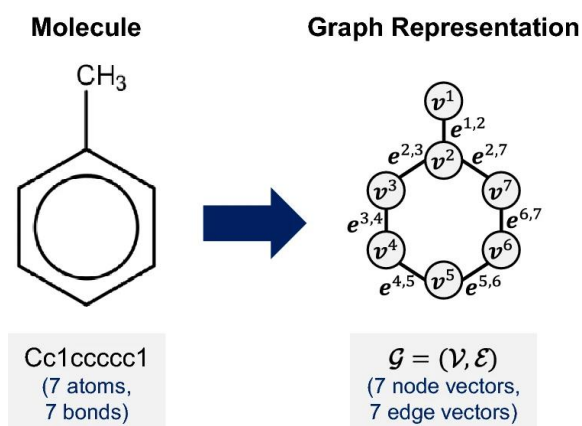
$$\theta = \{\theta_1, \theta_2, \theta_3, r_i, i = 1 \dots K\} \quad (2.40)$$

### 2.2.2 Graph Neural Networks

In this section, we discuss the more recent advances of machine learning in chemistry centred around graph neural networks. This approach is distinctly different from other methods previously mentioned and the graph representations of compounds used appear to be readily applicable to the structures seen in chemistry.

Instead of using descriptor values to encode a particular compound as seen in previous modelling methods described, graph neural networks take the molecule structure itself as a direct input (**Figure 2.4**).



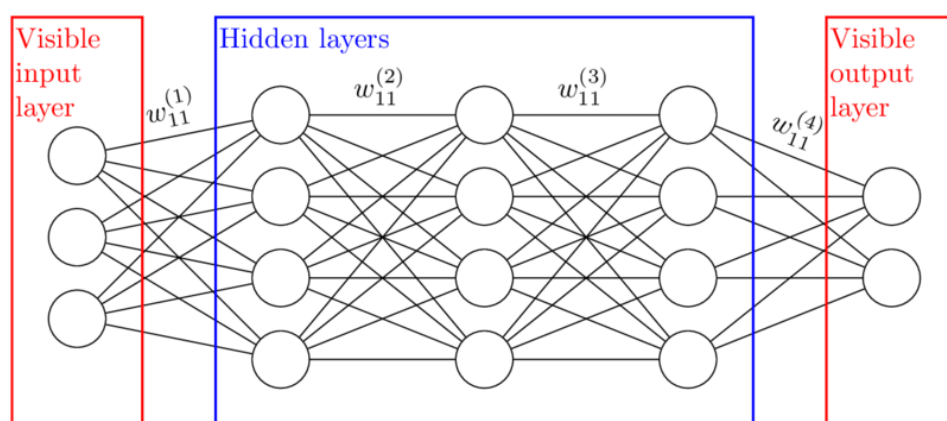


**Figure 2.4:** Graph representation example of methylbenzene. Figure reproduced from reference.<sup>198</sup>

Since graphs are built upon nodes and edges this is readily translated to molecules where atoms are represented as nodes and bonds are represented as edges. This structural representation allows a greater understanding of the relationship between the predicted value of interest and the structural connectivity that gives rise to the observed value.

Neural network architecture is inspired by the interconnectivity of the human brain (**Figure 2.5**) and is designed to gain an understanding of the interconnectivity between different features and the output value. The nodes or “neurons” in the network are typically split into layers, an input layer where the raw feature data is loaded into the network; hidden layers where a number of mathematical operations may be performed on the data to transform it and learn the non-linear relationships between features; and the output layer where the learned relationship is utilised to give the output value. During training, the weights and connectivity of each node in the hidden layers are altered to minimise a loss function. Seminal work on the neural network architecture by Hopfield in 1982 led to the development of the Hopfield

Network, a form of recurrent network with the ability to store and recall patterns in data. This was initially applied in image recognition tasks and its' memory retrieval abilities mimic the structure of the human brain.<sup>199</sup> This led to the award of the 2024 Nobel Prize in Physics for the foundational work that led to the advancements in the application of neural networks in machine learning.

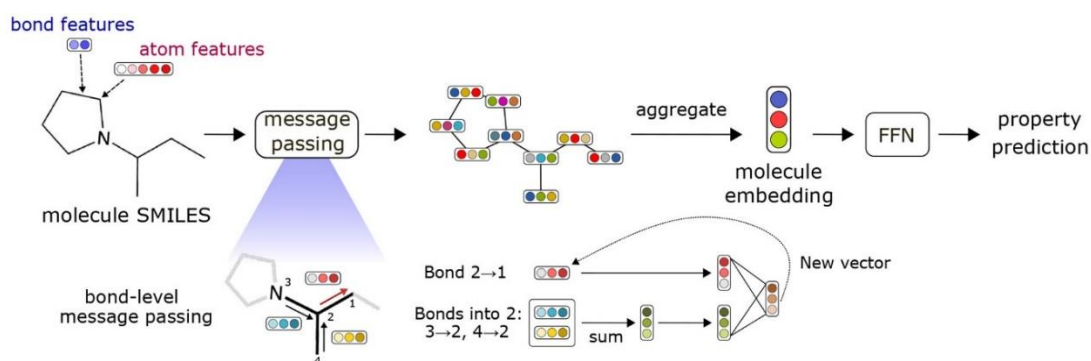


**Figure 2.5:** Basic neural network architecture.

Recent advances in neural networks<sup>200</sup> utilise a concatenation of different neural network architectures to provide greater predictive performance in the chemistry field, namely chemical property prediction. Alongside the use of a graph representation of each compound in the input, the graphs can be featurised further with QM descriptors that provide more information on the properties of each node in order to gain a greater understanding of the compound's activity. The use of a directed message parsing neural network in the ChemProp<sup>200</sup> modelling package serves to pass information about the atom/bond's local environment to generate local embeddings, giving the model a greater understanding of the other factors

influencing a particular atom's reactivity. These local embeddings are then combined to generate a molecular embedding where all these atom-bond relationships are used to infer the output variable (**Figure 2.6**). This neural network approach has shown great promise in molecular property prediction tasks given a large enough dataset size. Published applications of the use of this model architecture include the development of a novel family of antibiotic structures. In this work, structures of known antibiotic compounds were given to the model and trained upon these compounds' ability to inhibit the growth of *Staphylococcus aureus*, a key contributor to the concerns of antibiotic resistance, namely methicillin-resistant *S. aureus* (MRSA), as well as vancomycin-resistant *Enterococcus*.<sup>201</sup> This model was then used to predict the likelihood of inhibition of these bacteria on 283 novel compounds. Of these 283 compounds, one was found to show highly promising antibiotic activity across both of these strains of bacteria while avoiding resistance in *in-vitro* experiments. Another group used ChemProp to predict antibiotic activity against an array of different bacterial strains and found that one compound, halicin, was structurally divergent from other commercially used antibiotics and had activity against *Mycobacterium tuberculosis* and carbapenem-resistant *Enterobacteriaceae*.<sup>202</sup> Lastly, another group has successfully trained ChemProp on 41 different ADMET datasets to predict these important pharmacological metrics

faster and more accurately than any other tool previously developed, with one million compounds predicted in just 3.1 hours.<sup>203</sup>



**Figure 2.6:** ChemProp neural network architecture. Figure reproduced from reference.<sup>200</sup>

## 2.3 Related Work in Prediction of C-H Functionalisation Regiochemistry

Some exploratory work has been published on understanding and modelling this reaction in the past. Duan *et al.* recently published a paper detailing the behaviours of some of the carbon radical species typically found in this reaction, where the primarily used  $\cdot\text{CF}_3$  radical was said to display electrophilic properties, while the second most common  $\cdot\text{CF}_2\text{H}$  radical was found to display nucleophilic behaviour on select examples tested with both computational and experimental methods.<sup>204</sup> Since the differences in electrophilicity are subtle, atom-centred charges are insufficient to predict the correct site of reaction for a given substrate consistently and accurately. Instead, the interactions between the radical singly occupied molecular orbital and the substrate's highest occupied and lowest unoccupied molecular orbitals were

used, requiring a high level of computational cost (M06-2X/6-311+G(d,p)//SMD-M06-2X/def2-QZVPP). Since the level of computational cost is high for these calculations, especially on complex aromatic drug-like species, other methods of prediction warrant investigation.

Li *et al.* employed machine learning to predict the regiochemistry of this reaction,<sup>119</sup> focusing on relatively simple substituted heteroarenes. A random forest was trained on molecules represented by descriptors with a physical organic basis, such as bond orders, partial atomic charges and buried volume, computed from B3LYP/6-311+G(2d, p) calculations. The regioselectivities predicted by the random forest were compared to those computed from DFT free energy barriers (calculated at the M06-2x/def2-TZVP level) of the competing radical additions, with encouraging results; a site accuracy of 94% and a selectivity accuracy of 90% were achieved on out-of-sample test data. Nippa *et al.* employed a combination of automated nanomolar high-throughput experimentation, literature data and graph neural networks to predict Minisci-type chemistry.<sup>205</sup> This machine learning approach was used to explore the substrate landscape and led ultimately to the synthesis of 30 novel, functionally modified molecules. Hyek *et al.* utilised a graph neural network featurised with Fukui indices<sup>206</sup> and supplementing node featurisation with transfer learning from <sup>13</sup>C NMR shifts.<sup>207</sup> They saw promising results with an accuracy of 96% and an area under the receiver operator curve (AUROC) of 0.75. However, the task was predicting the atom-wise probability of functionalisation and labelling groups of sites as potentially labile rather than the prediction of the most reactive site.

In cases where there are insufficient experimental data and/or empirical sampling is prohibitively expensive, quantum chemical calculation, at an appropriate level of theory, offers a means of data generation in a rapid and cost-effective manner. Automated procedures for locating transition states enable calculations at a scale and with a generality which has not previously been readily achievable. One example is the program, autodE, which is an open-source tool that can locate transition states and minima, in order to provide a full reaction energy profile with minimal human intervention.<sup>208</sup> This is done by generating molecular graphs of each of the reactant and product structures and then finding the active bonds (or edges in the graph) that are changed between each structure. Then a series of constrained optimisations are performed along the path between reactants and products to locate the saddle point region, generating the transition state guess. This guess is fed into an unconstrained transition state search, using the optimisation algorithms in the quantum chemistry package of choice selected by the user. Another example, from Friesner *et al.*, utilised a similar workflow to automate transition state searches given known structures of separated reactants and products.<sup>209</sup> The combination of machine learning approaches and databases derived from quantum chemical calculations is proving to be powerful. For example, machine learning models have been used to predict the electronic energy and the free energy of small organic molecules, with a mean absolute error of 1.2 kcal mol<sup>-1</sup>.<sup>210</sup>

## 2.4 Application of Theory in this Work

In this research, we will utilise quantum chemical calculation at varying levels of theory to find a relationship between calculated properties and the experimentally observed regioselectivity for this C-H functionalisation reaction. With the chosen calculated property, we will then generate an artificial dataset of calculated regioselectivities for a wide array of compounds that will be used to build machine learning models to predict this regiochemistry on unseen examples.

# Chapter 3 - Regioselectivity

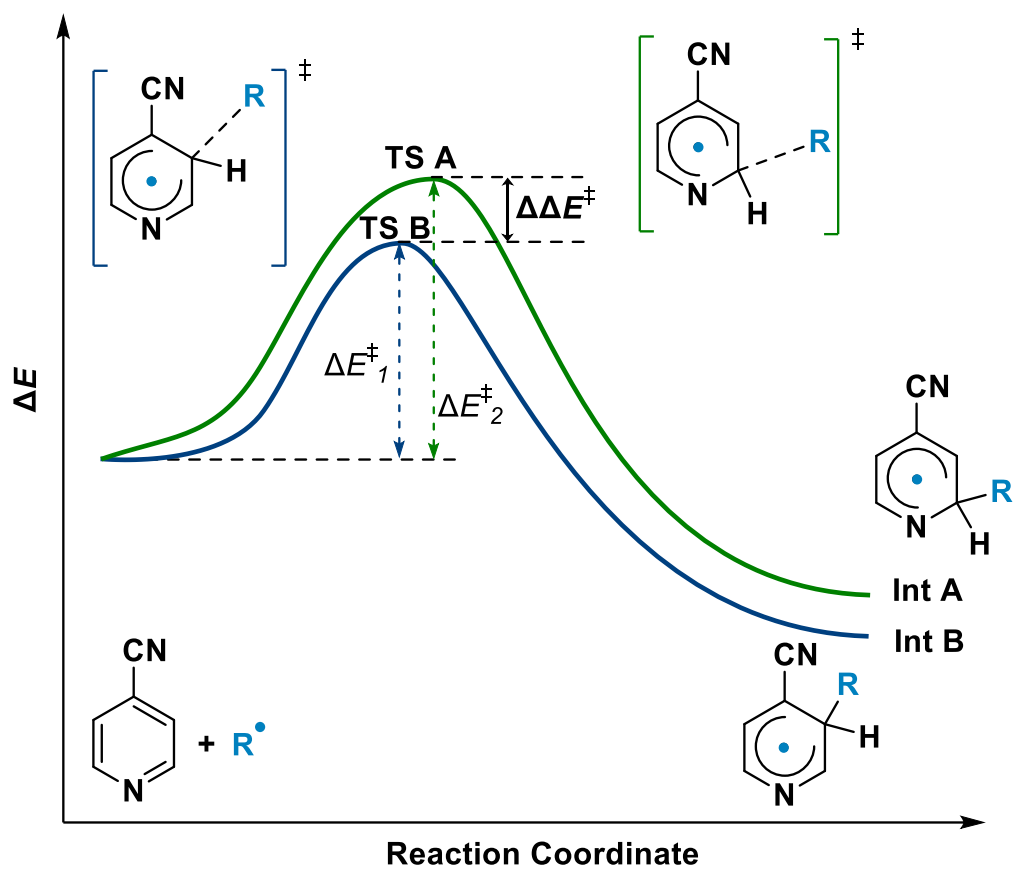
## Prediction

### 3.1 Introduction and Methods

In this chapter, we explore the different methods attempted for predicting regioselectivity. One approach is to investigate whether there are inherent features of the site of reaction that make it more susceptible to radical attack. This can be probed with the calculation of atom-centred charges on the substrate. In this method, a potential site of reaction may be more electron-rich or electron-deficient than other possible sites within the molecule, making it more susceptible to electrophilic/nucleophilic attack in its neutral state. Another possible method of study in this domain previously discussed is the analysis of the condensed Fukui function for the molecule through calculation of the Fukui indices at each potential site of reaction. In this method the site's propensity to gain/lose an electron is assessed rather than directly measuring the charge state of the nuclei alone. For the initial investigation, atom-centred charges were calculated for each compound in a benchmark set with experimentally validated regioselectivity information to understand whether there was an observable trend between the most likely site of reaction and that site's atomic charge.



Another approach is to understand the electronic and steric features of the transition state at each potential site of reaction that makes a particular site more conducive to radical attack. Examples of these features may be the diffuseness of the molecular orbitals of the substrate; understanding the susceptibility of a molecule's HOMO or LUMO orbitals to electrophilic/nucleophilic attack at a given site; or the conformational reorganisation needed to expose a particular site to radical attack. This can be probed by the location of the transition state and subsequent calculation of the activation energy of the addition of the carbon-centred radical to each potential site of the reaction. This method has been previously used to predict behaviour for a number of different reaction classes including nucleophilic aromatic photosubstitution.<sup>211</sup> Since the rate-limiting step for this C-H functionalisation is the addition of the carbon-centred radical to the heterocycle, the activation energy for this addition is calculated using quantum chemistry methods and the lowest activation energy site is compared with the experimentally observed site of reaction. The difference in activation energy between transition states at each potential site of reaction can then be used to estimate the regioisomeric ratio of products (Figure 3.1).



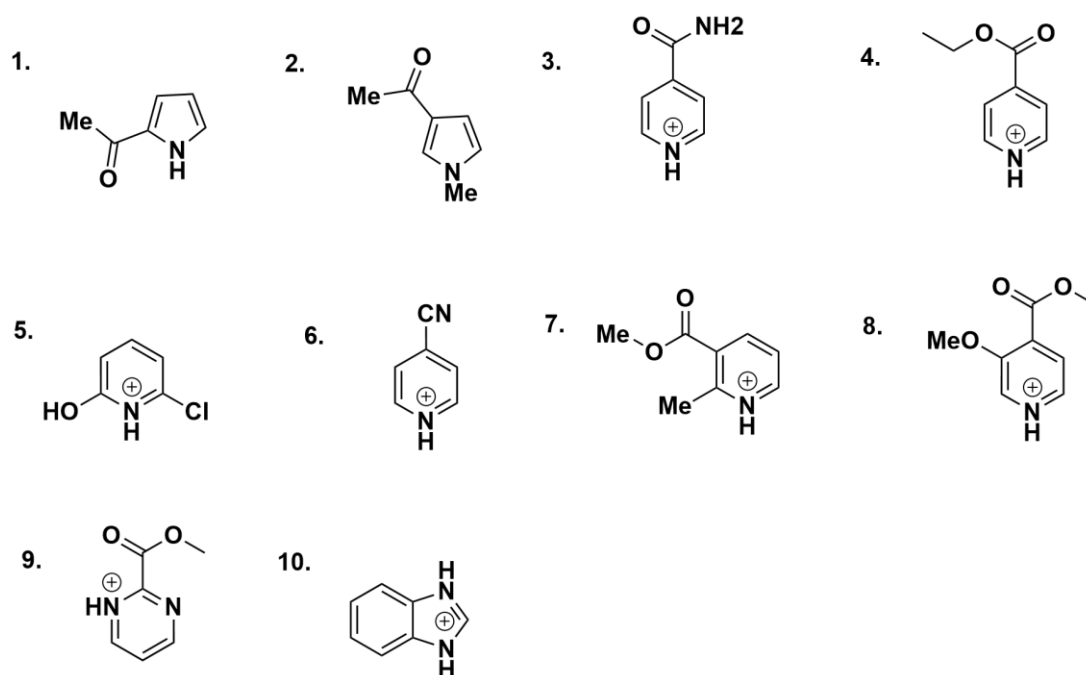
**Figure 3.1:** Site selectivity is determined by activation energy.

## 3.2 Results and Discussion

Below we explore the results of the investigations into these approaches applied to the sulfinate-mediated C-H functionalisation reaction of interest.

### 3.2.1 Atom-Centred Charges

The first approach attempted was calculating the atom-centred charges on one of the ten fragments (**Figure 3.2**) using the Gaussian 16 software package.<sup>212</sup> These compounds were gathered from a single source in literature<sup>119</sup> and were selected due to the increased confidence in experimental validity and their small size, enabling rapid calculation of these properties. In these calculations the geometry of the fragment was obtained from a M06-2X/def2-TZVP optimisation and single-point calculation of the structure. The predictive capabilities of two different methods were assessed, Hirshfeld charges and ADCH, which account for the deficiencies of Hirshfeld charges by including the atomic dipole moment of each atom in the calculation.<sup>181</sup> In the first experiment, both Hirshfeld charges and ADCH showed that the most electrophilic carbon atom (most positive carbon) was the site that was attacked by the radical in experiment. This result led to the calculation of the charges on the rest of the ten compounds in the set.



**Figure 3.2** Set of 10 compounds used for the initial regioselectivity investigation.

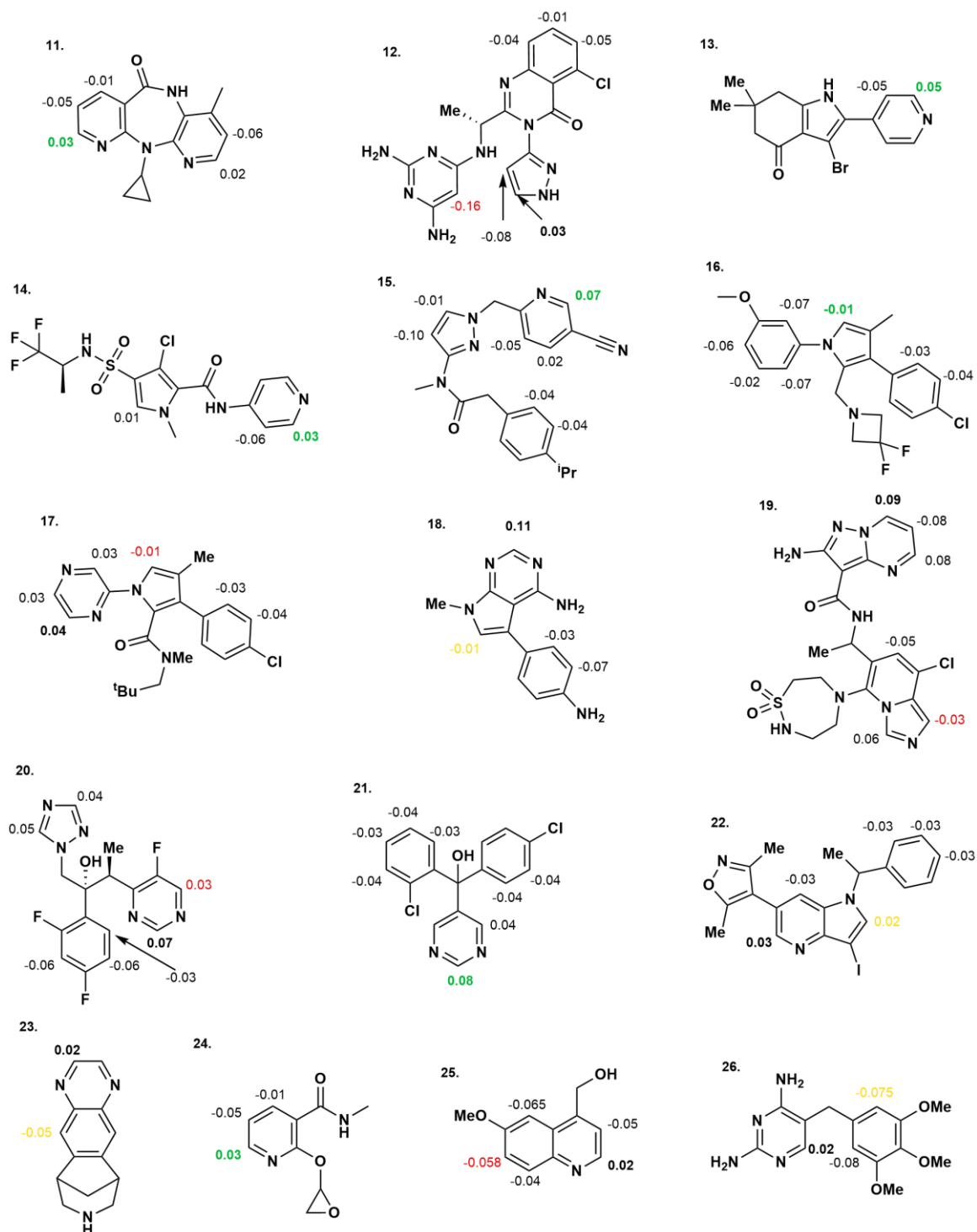
The preferred site of reaction was predicted based on the most positively charged carbon atom, using the Hirshfeld charges, and in an independent prediction, using ADCH charges. Both Hirshfeld charges and ADCH were able to predict the preferred site of reaction in the set, with ADCH predicting ten out of ten compounds correctly and nine out of ten predicted by Hirshfeld charges. The promising results of this preliminary investigation suggested that the charge descriptor should be tested further on drug-like compounds. To reduce the cost of the generation of these charges further, the geometry used to calculate the atom-centred charges was also optimised using Hartree-Fock with an M06-2X/def2-TZVP single point calculation, probing the dependency of the prediction on input geometry.

A set of 16 compounds (**Figure 3.3**) was taken from experimental papers<sup>213</sup> and the compounds were chosen based on a few criteria; the compound must have more than one possible site of reaction and the experimentally observed site must be known exactly (some compounds show substitution is isolated to a ring and not a particular site which is undesirable in this test). The results of this investigation firstly showed that the qualitative predicted site was invariant to HF/6-31G\* and M06-2X/def2-TZVP input geometries. Secondly, ADCH did not predict the correct site of reaction in drug-like molecules accurately. The observed site was commonly far from the most positively charged carbon in the compound and it seemed to represent the aromatic heterocycles present in these molecules as significantly polarised. Therefore, we do not consider ADCH further. However, Hirshfeld charges proved to be accurate in their prediction of site of reaction, with seven of the 16 compounds correctly predicted by this charge method. Of the remaining nine compounds, Hirshfeld charges predicted four sites that were only very slightly more positively charged than the experimental site which was the second most positive.

Incorrect predictions in this set can be separated into two categories, predicting a site within a different ring system to the experimental site and predicting a different site within the correct ring in the compound. Examples of the first error include compounds **12**, **17** and **19** and the second error is compound **20**. Charge predicted sites seem to be focused on the most nitrogen-rich heteroarenes and the preferred site is directly adjacent to a nitrogen atom in the ring. This behaviour is predictive of experiment in compounds with one heteroarene and one other aromatic ring in the system. However, it may fall short in compounds where two heteroarenes are found.

More testing was needed to probe whether charge predictions favour 5-membered or 6-membered heteroarenes when both instances are seen in a molecule and fully assess whether this charge method was predictive across a wide chemical space. However, considering this set of compounds contained a more diverse array of structures (compound **26**), and some structures contained multiple heteroarenes (compound **12**) with multiple viable sites of reaction, the results of this investigation were promising although this needed further evaluation on more drug-like compounds.

The applicability of Hirshfeld charge as a descriptor of regioselectivity prediction was also probed further by testing on a wider variety of compounds. To generate this new set of compounds, the review by Njardarson and co-workers<sup>214</sup> provided a list of the most used 5-membered and 6-membered nitrogen-containing heterocycles in drugs on the market. The compounds in this next set displayed atypical regioselective behaviour compared to the other compounds previously tested and therefore were used to assess the capability of Hirshfeld charges to predict the correct site of reaction in these more unusual cases.

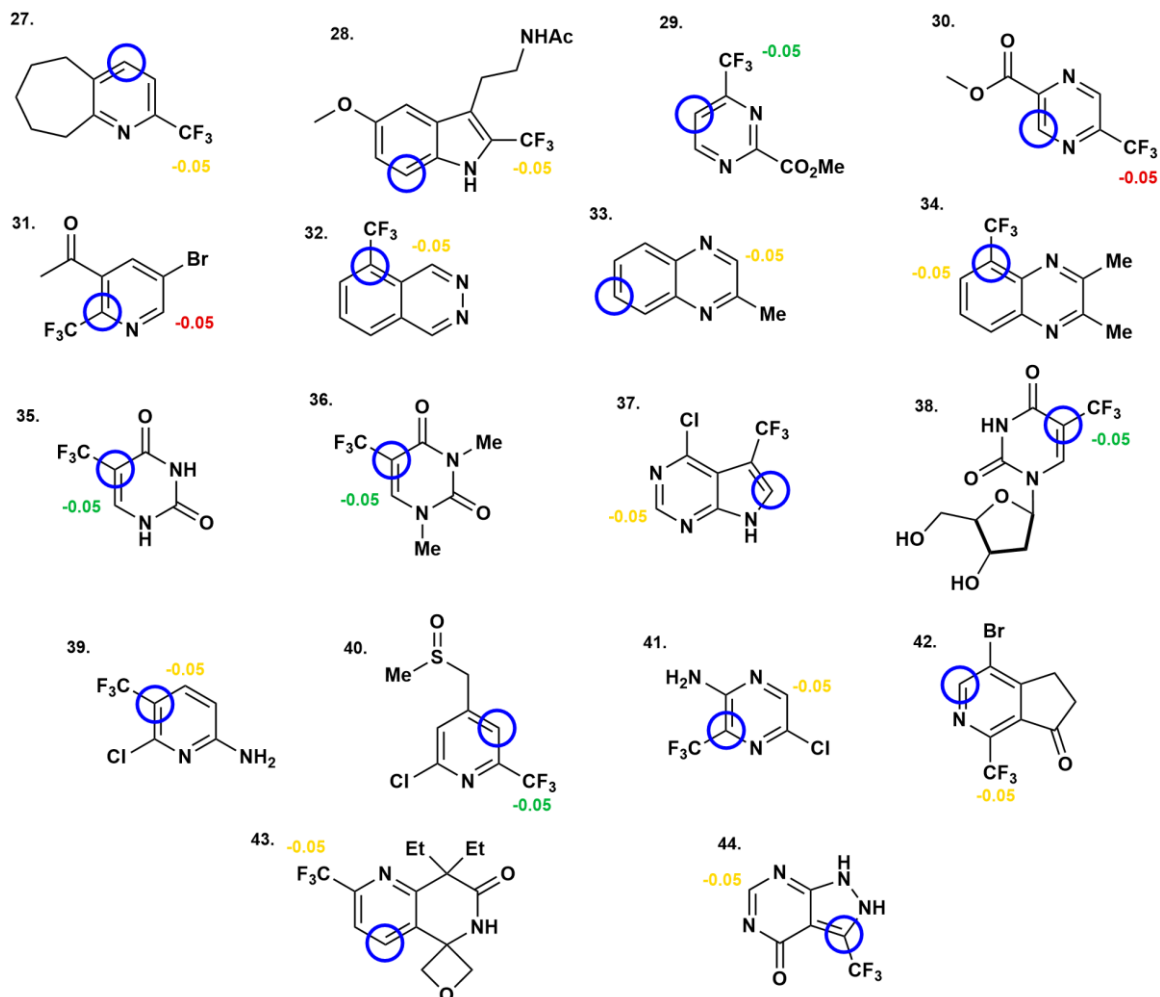


**Figure 3.3** Set of 16 drug-like compounds (a-p) used to evaluate Hirshfeld charge as a predictor of regioselectivity. Colour-coded sites are the positions where substitution occurs in experiment. Green sites are where the most electrophilic carbon atoms are the experimental sites of reaction, yellow is when the experimental site is the second most positive and red is when the experimental site is third most positive or lower. Sites in bold are the most positively charged sites and therefore the charge-predicted sites.

As shown in **Figure 3.4**, Hirshfeld charge was a poor descriptor for prediction of regioselectivity of this benchmark set. In many of these compounds, substitution does not occur at the position *ortho* to an aromatic nitrogen atom, unlike previous examples in other compound sets. Therefore, the atom-centred charge predictions which typically favour these *ortho* sites do not accurately predict experiment when other substitution patterns are observed. For example, **32** shows that when a carbon atom *ortho* to an aromatic nitrogen atom is available, Hirshfeld charges predict that substitution will occur in this position. However, experiment shows that substitution occurs in a different ring in this bicyclic system. Another example is compounds **35** and **36**, where Hirshfeld charge has a high level of confidence in its prediction that substitution occurs *ortho* to the nitrogen in the ring when experiment shows the site of reaction is adjacent to the carbonyl. Compound **37** shows other typical Hirshfeld prediction patterns that do not apply to this set. Other experimental sets show that when an aromatic carbon atom is available between two aromatic nitrogen atoms substitution occurs at this position. Hirshfeld charge depicts this site as noticeably more positively charged and so is the predicted site of reaction. However, in this example, we see that substitution does not occur between these two nitrogen atoms, nor *ortho* to the other nitrogen atom in the five-membered ring, but *meta* to the nitrogen atom in the 5-membered heterocycle. This means that both the most positively charged and second most positively charged sites in the compound were not the experimentally observed site of reaction which indicates that atom-centred charges in a compound are not the only contributing factor to the selectivity of this reaction. Another method of regioselectivity prediction was required, sensitive to



both typical *ortho* to aromatic nitrogen and atypical regioselective behaviours seen in experiment.



**Figure 3.4** Benchmark set of 18 compounds (a-r) used to evaluate Hirshfeld charge's regioselective predictions. The structures give the experimentally observed site of reaction and the charge predicted site is given by the position of the charge (e) on the compound. The number corresponds to the difference in charge between the first and second most positive sites, with colour coding indicating confidence. Green represents a difference of 0.1 or greater, yellow between 0.01 and 0.09 and red is 0.009 or below, showing very little confidence. The blue circle then represents the second most positively charged site in the compound.

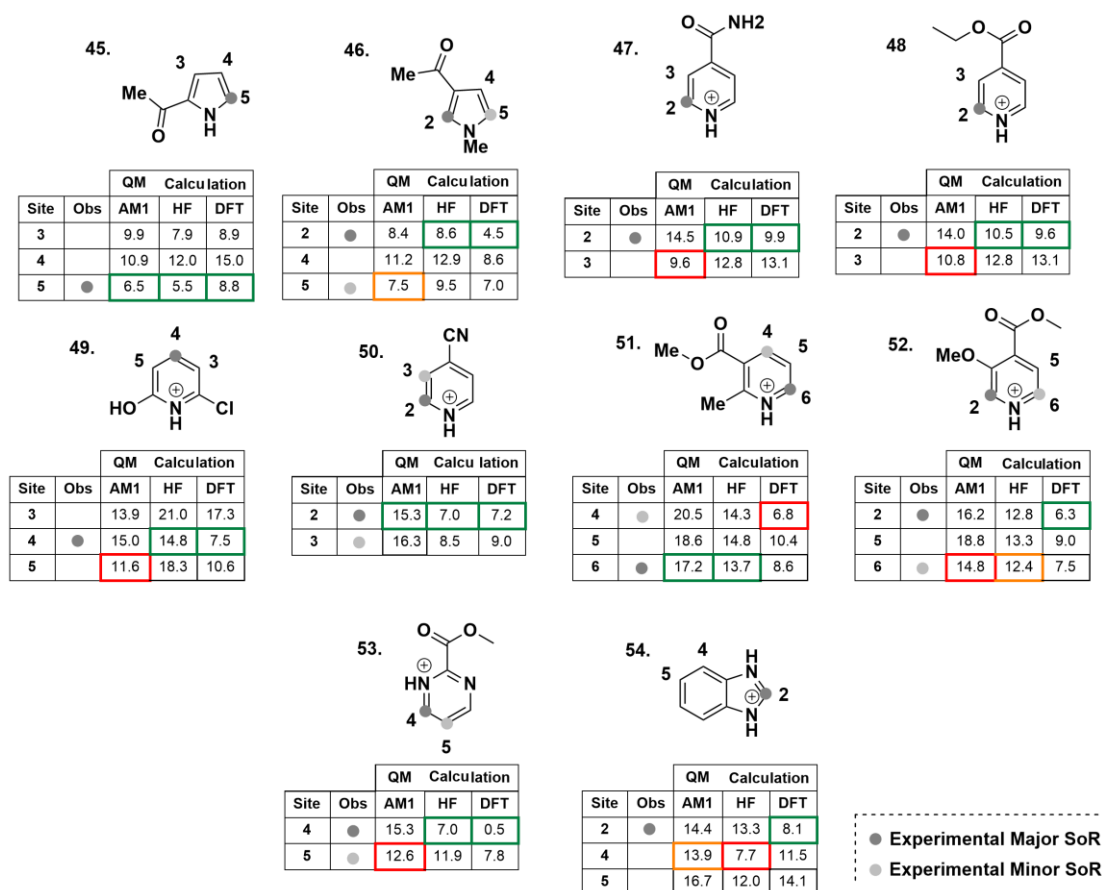
### 3.2.2 Activation Energy

Since it was apparent that more information about the site environment need be captured than the pure electronic information contained within atom-centred charges, the use of calculated activation energy as a descriptor to predict regioselectivity was evaluated. This descriptor focuses particularly on the rate-limiting step of the reaction and accounts for both the electronic environment of the site of interest but also the sterics of the surroundings and any conformational changes that may need to occur for the site of reaction to be sufficiently exposed to radical attack. This additional information may allow activation energy to be more capable of predicting other substitution patterns and more accurately replicate the experimental behaviour of this reaction compared to Hirshfeld charges.

We began with an assessment of the quantum chemical method that offers the best balance of accuracy and cost for our purposes, using a preliminary dataset, collated from the literature, for which there are experimental data on the position of substitution by one of three carbon radical species,  $\cdot\text{CF}_3$ ,  $\cdot\text{CF}_2\text{H}$  and  $\cdot\text{CF}_2\text{Me}$ . The dataset comprised 10 relatively small compounds and 26 sites of reaction (**Figure 3.5**). Activation energies were computed with the NWChem software package<sup>215</sup> at the AM1, Hartree-Fock (HF) and DFT (M06-2x/def2-TZVP) levels. These energies were used to calculate regioselectivity ratios and compared with experimentally observed regioselectivity. The functional and basis set for the DFT calculations were selected based on the findings of St. John *et al.*,<sup>216</sup> where the performance of many different functional and basis set combinations was assessed

in the energy calculations of radical organic species. In this study M06-2x/def2-TZVP was shown to give the smallest error with the least computational cost of many functional/basis set pairings when compared to much more costly coupled cluster calculations in the simulation of homolytic carbon bond formation, and so this functional/basis set was selected for our investigative work.

**Figure 3.5** shows the predictions of each method on the 10 compounds. AM1 shows poor predictive ability especially on protonated pyridine and pyrimidine species, with the difference between the lowest activation energy site and experimental site commonly being greater than 2 kcal mol<sup>-1</sup>. Hartree-Fock showed promising results in regioselectivity prediction, with eight out of the ten compounds correctly predicted. Of the two incorrect predictions, compound **52** was incorrect by 0.4 kcal mol<sup>-1</sup> between the lowest energy site and experimental observation. This prediction, based on a ratio close to 1:1 for the lowest activation energy site and second lowest site, is understandable, as multiple products are observed in experiment. The other incorrect prediction of compound **54** occurs for a moiety which is an outlier in the set. However, more testing was needed to understand whether this prediction was anomalous or whether Hartree-Fock was unable to predict these bicyclic heteroarene systems.

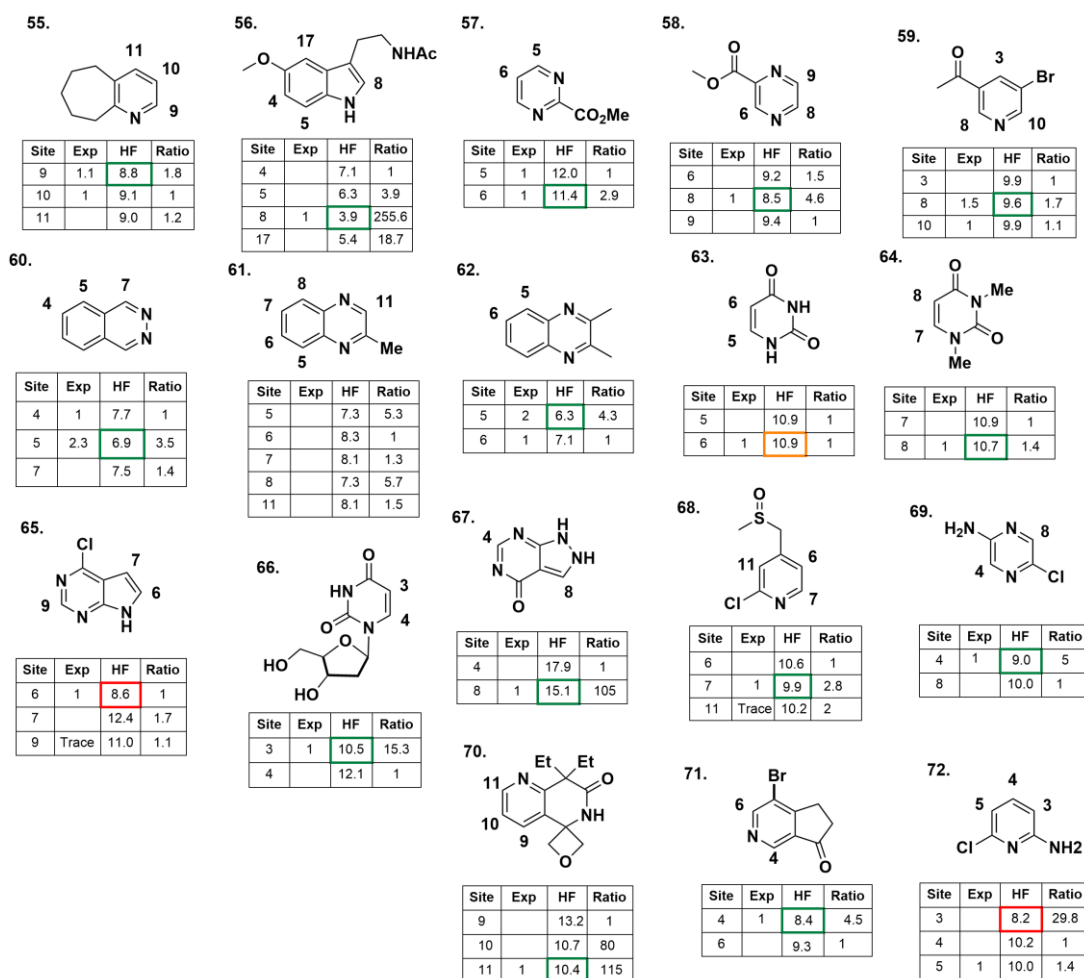


**Figure 3.5** Activation energies (kcal mol<sup>-1</sup>) for each site in the preliminary set of ten compounds. Experimental regioselectivities are given by grey dots on the structure. Colour coding corresponds to how well the method agrees with experimental observation (given by the Obs column), where green boxes show the method agrees with experimentally observed major product, yellow boxes are when the experimentally observed major product is within 1 kcal mol<sup>-1</sup> of the lowest activation energy for that method. Red boxes are when the prediction is more than 1 kcal mol<sup>-1</sup> different from the observed major product.

DFT (M06-2x/def2-TZVP) showed the greatest accuracy in the prediction of regioselectivity of the C–H functionalisation reaction, with nine out of ten compounds correctly predicted. However, the computational cost of this approach is considerable, due to the multiple rounds of transition state searches required to obtain the transition state geometry. While convergence was slightly faster in the substituted pyrroles in Figure 3.5, the cost was still much greater than Hartree-Fock and the latter also gave the same correct qualitative prediction. In the M06-2x functional, the Hartree-Fock energy makes up 54% of the exchange energy.<sup>217</sup> This

explains the similarity in predictive performance between the DFT and Hartree-Fock methods and the slight boost in performance over the already successful HF predictions, albeit at a significantly increased expense.

After this preliminary investigation, our assessment of the performance of HF/6-31G\* was expanded to a larger set of more drug-like compounds, since these are the target for a final model. The compounds gathered from literature experimental data had greater complexity than those previously tested with more diverse steric and electronic properties around each potential site of reaction. **Figure 3.6** shows the HF/6-31G\* predicted regioselectivity well, agreeing with the literature reported qualitative site of reaction. Importantly, we see that compound **64** has a greater selectivity than compound **63**, implying that the steric information on site availability is captured in this calculation. This is due to the additional steric hindrance at site 7 in compound **64** provided by the adjacent N-methyl group compared to the N-H group adjacent to site 5 in compound **63**.



**Figure 3.6** Activation energies (kcal mol<sup>-1</sup>) for each potential site of reaction in the benchmark set and the experimentally observed ratios of each product. Green boxes indicate an agreement with experiment, yellow indicates a disagreement between activation energy and experiment but the difference between the lowest activation energy and correct site is within 1 kcal mol<sup>-1</sup> and red when the difference between the lowest activation energy and experimentally observed product is greater than 1 kcal mol<sup>-1</sup>. Regioisomeric ratio calculated from the difference in activation energy between sites is also given.

When extended further to a larger set of drug-like compounds, HF/6-31G\* calculations proved to be predictive of experiment, achieving a 96% correct classification of the most likely site prediction for 23 drug-like compounds comprising 68 sites. However, this level of accuracy was only achieved after re-performing some of the original syntheses in-house, for cases which initially appeared to be outliers,

using a consistent set of reaction conditions, as employed by Baran and co-workers.<sup>140</sup>

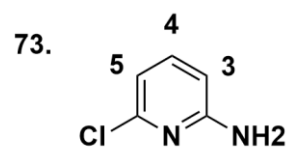
Important aspects of the literature experimental data include the conditions, *e.g.*, solvent selection, reagent concentration, temperature, and pH, under which the reactions were performed, and the method used to determine the regioisomeric ratio. An acidic environment may alter the protonation state of the substrate. Since there are multiple sites of protonation, particularly in drug-like molecules, pH can influence the regioselectivity. Differences in solvent may change the nature of interactions between solvent and substrate, leading to increased or decreased lability of a particular site. In our experimental work, the ratio of regioisomers was determined using quantitative <sup>19</sup>F NMR yields on the crude mixture and, where no literature spectroscopic data were available for a given regioisomer, purification via silica gel column chromatography was carried out to assign each isomer to the corresponding <sup>19</sup>F NMR signal.

To evaluate the efficacy of activation energy as a means of predicting experiment, it is imperative that good quality data is used and is even more critical when using this data to train a machine learning model. Thus, synthetic work was carried out on the outlier compounds (see section 7.2 for details) in the previous set to understand whether the (few) disagreements between our HF/6-31G\* calculations and reported experimental data were due to a failure in the predictive ability of the activation energy method or due to one of the other factors mentioned above. Reaction conditions for the study were based on the original Baran papers,<sup>142</sup> though since

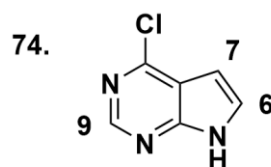
mono-substitution is the focus of this research a second portion of the zinc sulfinate salt was not added if starting material remained after 24h. This removed the possibility of di- or poly-substitution, which would have otherwise precluded the ability to probe the innate regioselectivity of the unsubstituted compound.

The first experiment was conducted on compound **72** following the above procedure rather than the acidic conditions used in previous literature. **73** shows that while previously substitution was reported to occur solely at site 5, under standard conditions it occurs at site 3. While the quantitative ratio prediction is not correct the qualitative depiction of the preferred site of reaction now agrees with calculation. This continues in **74**, where previous literature reports functionalisation at site 7, whereas calculation predicts regioselectivity for the reaction at site 6 then site 9 followed by site 7. After performing this reaction under standard conditions, we see agreement between experiment and calculation, where not only was the preferred site predicted correctly, but the next most favourable site was also predicted with trace amounts of 9-functionalised product seen in the crude mixture. Since this reaction shows strong dependence on reaction conditions, any machine learning models generated would predict a compound's selectivity rather than reactivity with an understanding that the regioselectivity prediction is valid when performing the reaction under standard conditions.





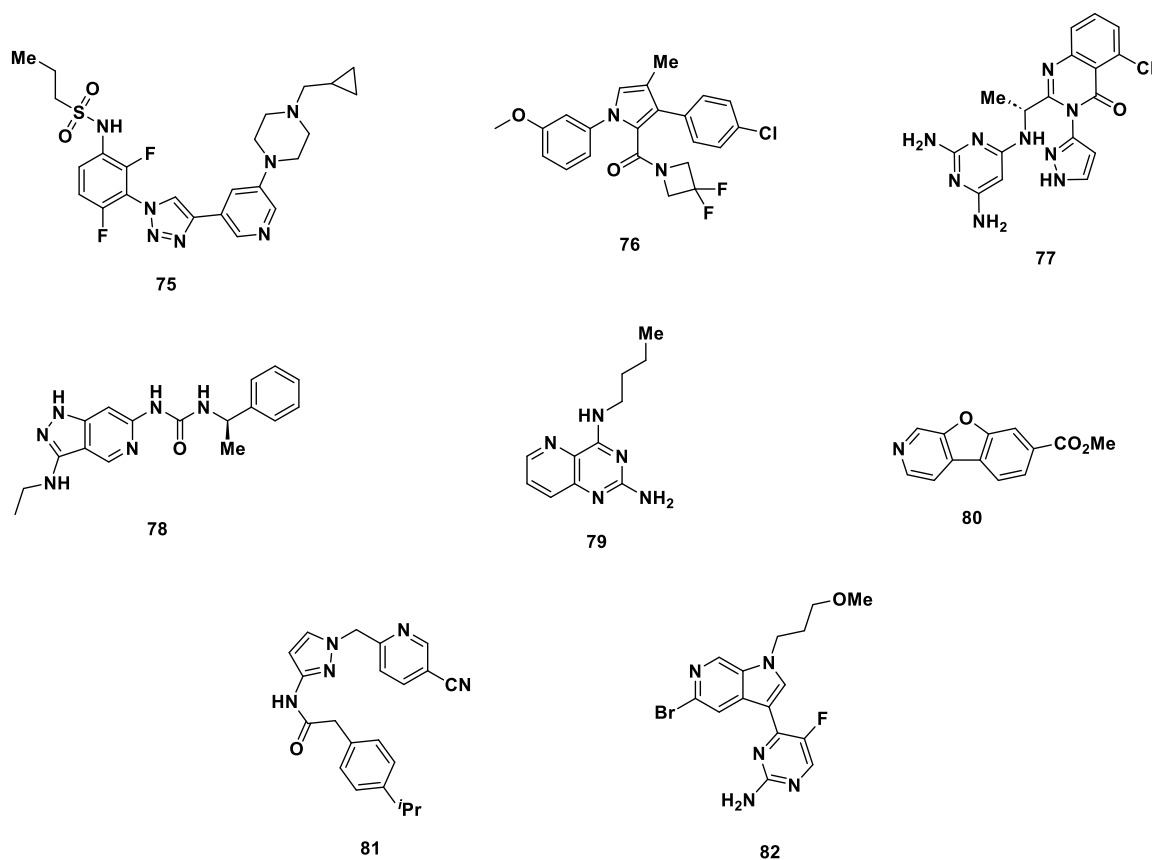
Site	Prev Exp	HF	HF Ratio	Our Exp
3		8.2	1	1
4		10.2	0.03	
5	1	10.0	0.05	0.25



Site	Prev Exp	HF	HF Ratio	Our Exp
6		8.6	1	1
7	1	12.4	<0.01	
9		11.0	0.02	Trace

**Figure 3.7** Experimental regioselectivities of compounds that previously disagreed with calculation. The calculated activation energies and regioisomeric ratio are shown alongside the new experimentally derived ratio of regioisomers. The green shading indicates an agreement between experimentally observed and calculated ratio of products.

### 3.3 Activation Energy Calculations on other Drug-like Compounds



**Figure 3.8** Other drug-like compounds where activation energy was compared with experiment.

When investigating the viability of activation energy as a predictor of regioselectivity on a series of drug-like compounds, there were other compounds tested that QM-derived activation energies did not correctly predict. **Figure 3.8** lists the compounds that did not agree with calculation. The reasons for the disagreements between calculation and experiment will be explained below.

Compound **75** failed to converge to a transition state on multiple attempts on each site of reaction. Compound **76** according to the literature was performed with sodium trifluoromethanesulfinate rather than zinc trifluoromethanesulfinate. The solvent system selected was also dimethylformamide rather than the biphasic mixture of dichloromethane and water used in the original discovery reaction methodology. Lastly, instead of TBHP, hydrogen peroxide was used in conjunction with  $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ . These conditions are noticeably varied from the original reaction scheme and any number of these changes could have a marked effect on regioselectivity prediction. Compound **77** was also performed using the sodium trifluoromethanesulfinate salt. This altered counterion could stabilise the intermediate of a different site of reaction and so give a different functionalised product compared to calculation. Compound **78** was performed using dichloroethane, water and dimethylsulfoxide alongside the use of the sodium derivative of the sulfinate salt. The solvent conditions in particular are crucially important in predicting regioselectivity with HF/6-31G\* transition states. In compound **79**, the reaction was performed with the addition of trifluoroacetic acid. Since there are many protonation sites within the molecule and it is difficult to deduce the correct site of protonation in the reaction, this protonation was not accounted for in the calculation. As such, regioselectivity prediction differed from experiment as it was proven that this dramatically affects regiochemistry. Compound **80** was performed using the difluoromethanesulfinate salt rather than the trifluoromethanesulfinate diversinate typically used in other calculations. This change in the structure of the radical species in the rate-determining step can

dramatically influence the regiochemistry of the reaction. In this example, the modification of the  $\text{CF}_3$  group to the  $\text{CF}_2\text{H}$  group means the radical is now more nucleophilic and so the preferred site of reaction is different to the more electrophilic  $\cdot\text{CF}_3$  radical. Similarly, **81** was also performed with the difluoromethanesulfinate salt rather than the trifluoromethanesulfinate. This was not the radical species that was used in the calculation so regioselectivity predictions will differ from experiment. Lastly, compound **82** was also performed using the  $\text{Zn}(\text{SO}_2\text{CF}_2\text{H})_2$  salt and so regiochemical behaviour was different to the predicted output of the calculation where the  $\cdot\text{CF}_3$  radical was used. In future, the validity of activation energy may be assessed on these compounds with differing reaction conditions with the use of solvent correction models in the calculation. This may allow for a greater understanding of the regioselective behaviour of substrates in different solvent conditions through a site's interaction with the solvent, making that site more or less labile when compared to standard conditions.

### 3.4 Conclusions

The results of this investigation show that activation energies are more predictive than Hirshfeld charges. Most of the benchmark set compounds' experimentally observed sites are the sites with the lowest activation energy, including those with atypical substitution patterns. This motivated efforts to build a machine learning model to predict regioselectivity using activation energy as the predictive method of

choice. Since the volume of experimental data is insufficient and the quality of data is too poor to use for machine learning, an artificial dataset must be generated. However, since the activation energy calculations for the above drug-like compounds manually took in excess of two weeks, an automation workflow was built to efficiently generate a dataset of sufficient size for modelling.

# Chapter 4 - Automation & Software

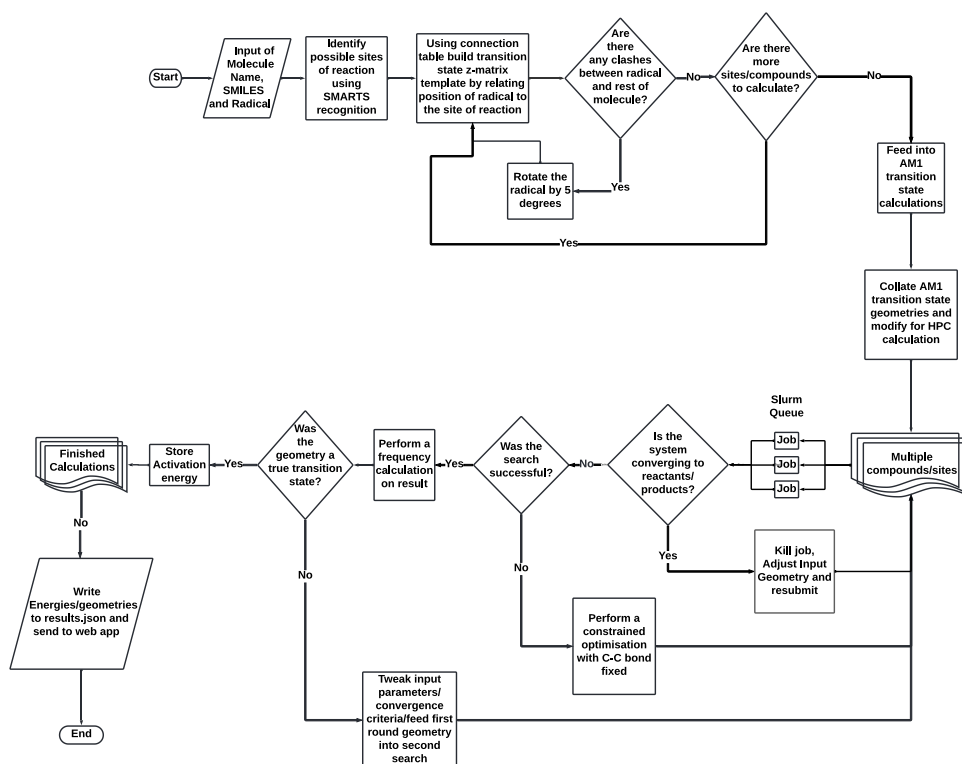
## Development

In order to efficiently perform the many thousands of transition state calculations required for the artificial dataset, a workflow was developed to handle and monitor these calculations running simultaneously on high-performance compute (HPC) facilities. The program known as Rega was created to utilise well-known quantum chemistry software packages to perform and monitor the transition state calculations required at scale in an autonomous fashion. This dramatically reduces the manual intervention required by the user and reduces the labour of the generation of activation energy-based datasets substantially. In this chapter, we discuss the steps employed by the Rega program in detail, outlining the procedures for the creation of pseudo-transition state geometries, their feeding into quantum chemistry packages for transition state search calculations and the monitoring of said calculations at scale.

### 4.1 Local Semi-empirical Calculations

In this section of the chapter, we will discuss the steps required to perform the initial local semi-empirical calculations that give a large reduction in cost in locating the final HF/6-31G\* transition states. While the regioselectivity predictions made by AM1 were previously shown to not be reflective of experiment in chapter 3, the dramatic

reduction in computational cost compared to *ab-initio* methods was attractive for the generation of transition state conformations close to the more accurate Hartree-Fock calculations. Therefore, semi-empirical calculations were utilised to perform the initial transition state search before being fed into the HF/6-31G\* calculations. This serves to reduce the number of SCF cycles required by the more costly calculation methods, increasing computational efficiency. The flow diagram shown in **Figure 4.1** outlines the main workflow of the software developed known as Rega which of which the first half will be explained in this section.



**Figure 4.1** A flow diagram outlining the full functionality of Rega.

### 4.1.1 Transition State Templates

In order to calculate HF/6-31G\* transition states rapidly for the prediction of regioselectivity, a variety of methods must be employed. Firstly, since the rate-limiting step of the reaction is known and the radical species is conserved, a pseudo-transition state can be constructed from a template and used for primary transition state searches. The user can specify which radical species they want to add to the program input, where a different pseudo-transition state will be built depending on the reaction being studied. The current groups that can be added are the  $\cdot\text{CF}_3$ ,  $\cdot\text{CF}_2\text{H}$  and  $\cdot i\text{Pr}$  radicals. Once this radical is selected the pseudo transition state is built.

To start, Open Babel<sup>218</sup> is used to generate a three-dimensional geometry of the substrate of interest in the form of Cartesian coordinates, where each atom's position is given in three-dimensional space. Since this initial output structure may have an irregular conformation, a cheap molecular mechanics optimisation (MMFF94)<sup>219</sup> is performed on the structure to find a lower energy conformation with which to start the pseudo transition state template-building process. To add the radical species close to the site of interest, the atom number must be found. To do this, the RDKit<sup>220</sup> package was used with SMARTS pattern recognition<sup>221</sup> to identify the aromatic carbon atoms that have only hydrogen as the other group connected to it. From this, a list of site numbers is generated with which to use for transition state calculations on each possible site of reaction. With this list and optimised structure, the radical species must be added to the system in a position close to the site of interest. Firstly, a directory is created for each potential site of reaction and the



Cartesian Coordinate .xyz file for the MMFF94 optimised reagent is copied into it to use as a starting point for the building of the pseudo transition state geometry. The Cartesian coordinate representation of the substrate geometry must be converted into a z-matrix, a representation where an atom's position in space is defined by its' distance, angle and dihedral angle to another atom in the system (see Chapter **7.4** for more detail). This conversion is completed using Open Babel and the atom number assignment is conserved between Cartesian and z-matrix representations. There are many possible forms that the z-matrix can take. The atoms chosen to relate the position of the newly added atom drastically impact the stability of the subsequent semi-empirical calculation. Therefore, the z-matrix form must be chosen carefully particularly when manually adding atoms to previously generated z-matrices. To add the new radical species to the system close to the site of interest in a stable manner, a connection table must be generated that relates each heavy atom in the carbon backbone to one another, and the other functional groups in the compound's position related to the atom bonded to it. This gives a more chemically relevant representation of the molecule as the position in space of a newly added atom is not associated with an atom many bond lengths away. The generation of this connection table was done using the ChemCoord Python package,<sup>222</sup> which was unique among other packages screened in testing where the carbon backbone was properly defined rather than each atom's position related to the first. Since ChemCoord takes the RDKit Cartesian Coordinate geometry as an input, atom number assignment is conserved in the connection table. This table was used to locate the site of interest and the surrounding atoms with which to relate the angle

and dihedral angle of the radical species. Importantly, the atoms surrounding the site of interest were chosen to be the heavy atoms in the backbone of the compound rather than the other substituents connected to the site of interest, namely the hydrogen present on the site of interest for the “angle atom”. The other heavy atoms two bond lengths away from the site of interest are used for the “dihedral atom”. Once these bond, angle and dihedral angle atom numbers have been defined the radical species is added to the z-matrix using predefined distances and angles taken from preliminary manual transition state searches from previous investigations. Once this modified z-matrix is generated it is then checked for any clashes between the newly added radical and any other atoms in the substrate. This is done by converting the newly generated z-matrix into a Cartesian Coordinate .xyz file using RDKit and the distance between each atom in space is calculated using 3D Pythagoras. If any of the radical atoms are within 1 Å from any other atom in the system, then the dihedral angle of the radical species is rotated 5° before checking again. This threshold of 1 Å was chosen since the subsequent semi-empirical calculations can adjust the conformation of the compound to accommodate this new radical species at this distance. If the added radical was closer to the substrate, it would cause the calculation to fail.

#### 4.1.2 Semi-empirical Calculations

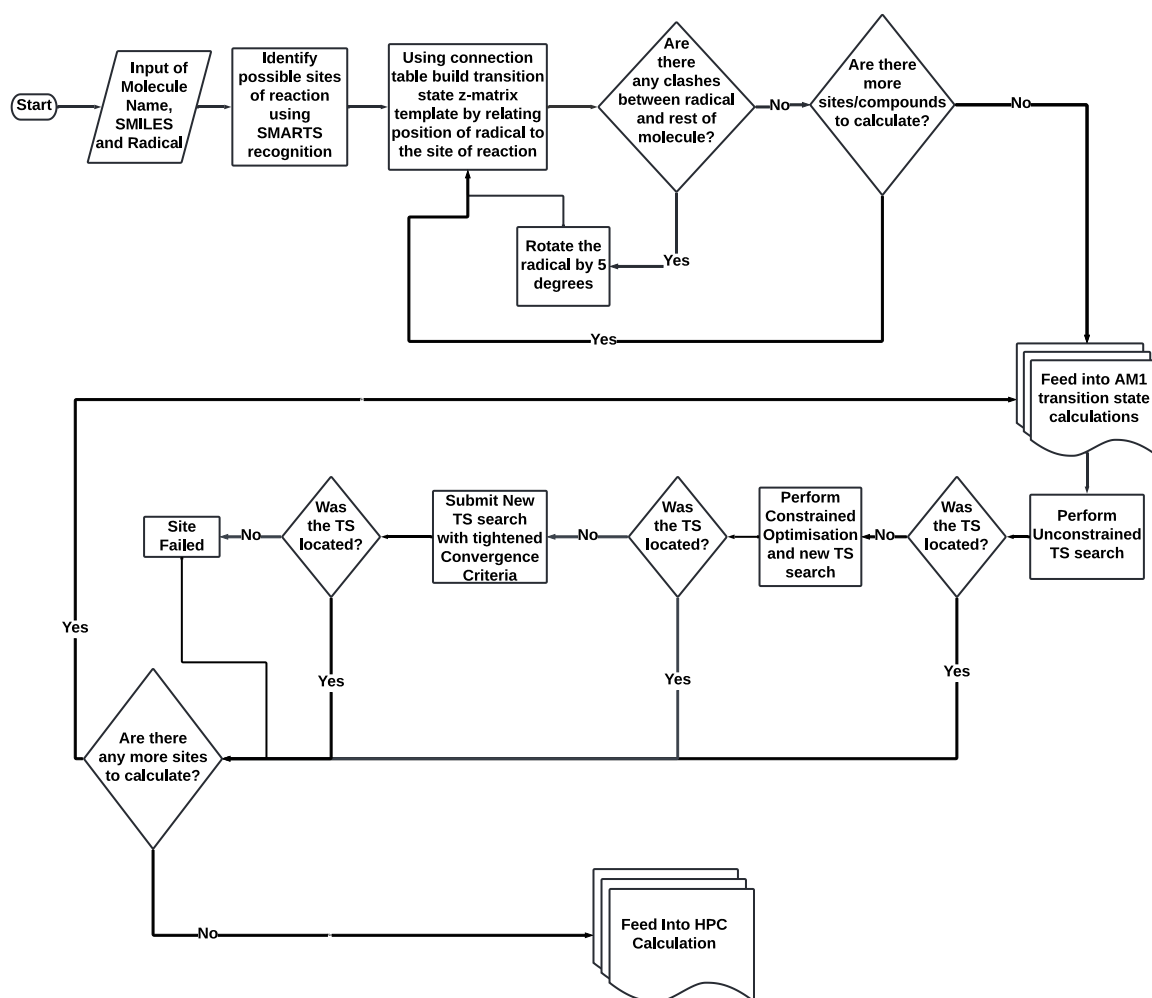
These new pseudo-transition state geometries are now used as input for a variety of calculations. All semi-empirical calculations were performed using the

MOPAC2016<sup>223</sup> software package and the AM1 method. Firstly, one site is selected to use an alternative template for reagent optimisation. The pseudo transition state geometry for this site is copied into a “reagent” directory before modification for the calculation of ground-state energy. In this “reagent” template all angles and dihedral angles are conserved but the “forming” C-C bond length between the attacking radical and the site of reaction is extended to 12 Å and the system is then allowed to relax using a geometry optimisation. Each reagent optimisation is then checked to see whether the calculation is completed successfully before moving on to the transition state searches. If the first reactant optimisation is not successful, another site of reaction is chosen to calculate the ground-state energy from.

Each potential site of reaction first undergoes an unconstrained transition state search using the pseudo transition state template built earlier. The result of that calculation is checked for completion and if finished, the resultant geometry undergoes distance checks to ensure the forming C-C bond is between an expected range that is typical for the AM1 transition state on these systems.

If the unconstrained first transition state search is unsuccessful, a constrained optimisation is performed with the site of reaction and carbon radical’s position fixed. This allows the rest of the substrate to alter its conformation to expose the site of reaction and make it more favourable to radical attack. This resultant geometry is checked and sent into an unconstrained transition state search. Any transition state searches that finish and pass the distance check function are then fed into a frequency calculation to ensure the true saddle point on the potential energy surface

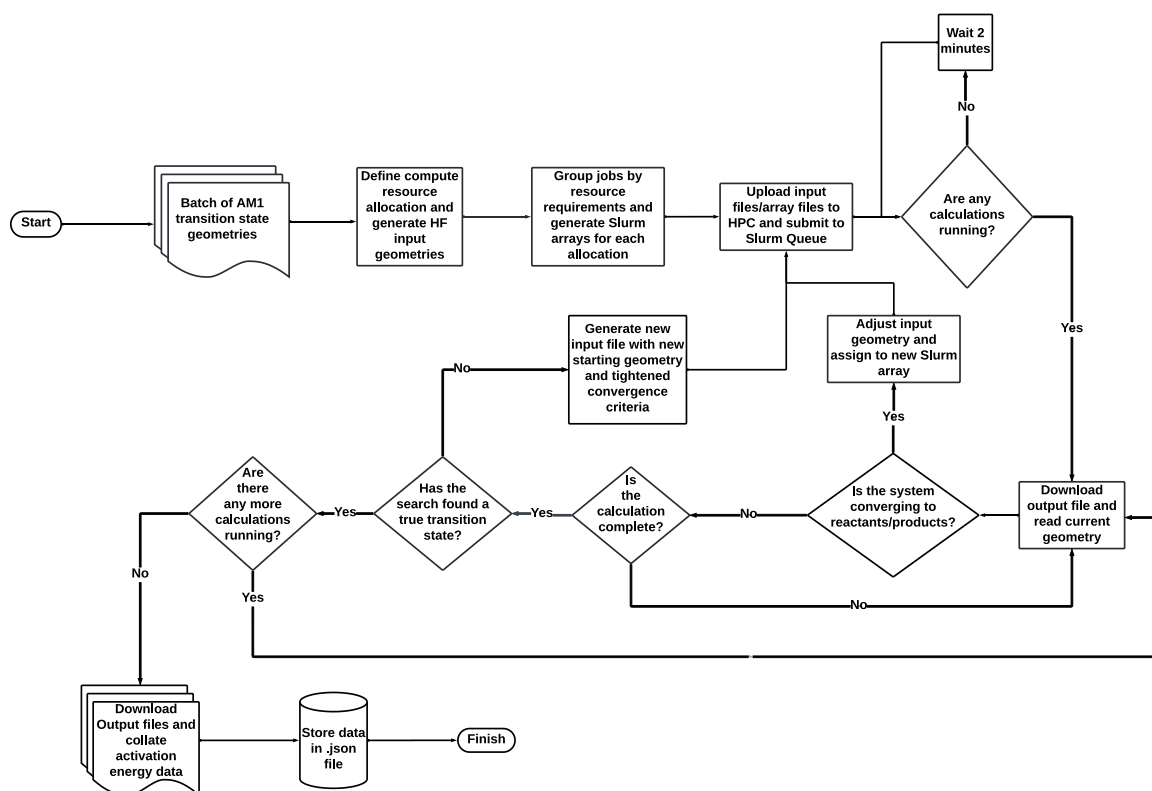
has been located. If there is a single imaginary frequency within an expected wavenumber range the transition state search is deemed successful and Rega moves on to the next site in the substrate. If there is more than one imaginary frequency but the first mode is within the typical range, a second transition state search is performed on this resultant geometry with tightened convergence criteria to remove spurious frequencies and converge on the true saddle point. The new semi-empirical transition state geometry is converted to a Cartesian Coordinate .xyz file for use in the subsequent Hartree-Fock calculations on the High-Performance Compute (HPC) facility. The semi-empirical (AM1) calculation workflow is summarised in **Figure 4.2**.



**Figure 4.2** Workflow for the semi-empirical calculation portion of Rega.

## 4.2 HPC Calculation

In this section we discuss the workflow to take the pre-calculated AM1 transition states as the starting point for the more costly HF/6-31G\* calculation using the High-performance compute (HPC) service of choice. The workflow for this portion of the Rega program is summarised in **Figure 4.3**.



**Figure 4.3** HPC workflow of Rega.

### 4.2.1 Modification of AM1 transition state

Since the semi-empirical transition state geometry systematically underestimates the carbon-carbon bond length of the attacking radical species, some modification of the output geometry must be done to make it suitable as a starting point for HF/6-31G\* calculation. Although this bond length needs to be modified, the conformation of the rest of the substrate is suitable for low-level calculation. The use of the modified semi-empirical transition state structure as a starting point for HF calculation serves to reduce the number of SCF cycles required to find the transition state in this more expensive method. Firstly, much like the generation of the pseudo

transition state template, the AM1 output geometry is converted to a z-matrix and with conserved atom mappings, the bond length of the added radical species is elongated from the typical 1.9-2.0 Å range seen in AM1 to 2.12 Å. This elongated bond is more typical of the bond lengths seen in the HF/6-31G\* transition states and the modification of the AM1 geometry ensures the first Hartree-Fock (HF) transition state search does not simply converge to the “product minima” on the potential energy surface.

#### 4.2.2 Generation of HPC input file and HPC resource allocation

The newly modified z-matrix is then converted back into Cartesian coordinates for use in the QM HPC calculation. Then, depending on the toggle switch chosen by the user in Rega, an input file is generated for either NWChem<sup>215</sup> or Gaussian16<sup>224</sup> with suitable keywords for the first HF transition state search. As part of this input file, the appropriate computational resource is required to be specified. To do this, the size of the system being calculated is noted by the number of heavy atoms. This number is then binned into one of seven core count groups, ranging from 8 to 32 cores per calculation. This is done to maximise computational efficiency, giving greater core allocation to larger systems that benefit from more computational resources per SCF cycle and smaller core counts to smaller systems that can see a reduction in efficiency when too many cores are given to a calculation. It is assumed that the reduction in efficiency in these cases is due to the communication between cores required to

complete the SCF cycle taking longer than the actual integrals that need to be calculated.

Each core allocation is also paired with a total memory allocation that is dependent on the number of cores. Memory per core is conserved between system size since we do not see a reduction in performance if not all memory allocated is used but larger core counts require more total memory to prevent crashing. Once these input files are generated, they are uploaded to the HPC ready for calculation. Semi-empirical calculation and subsequent input file generation is typically completed within one minute for each compound.

Alongside the transition state search calculations and ground-state optimisations required for the activation energy descriptor, other atom-specific properties such as electron density, electrostatic potential and Fukui indices<sup>225</sup> are calculated in a separate directory that can be used as other descriptors in the generation of machine learning models.

### 4.2.3 Job Submission and Monitoring

In order to submit calculations to an HPC service, users' jobs must be scheduled according to number of jobs and resources requested. This prevents smaller requirement jobs from being "stuck" behind more resource-intensive jobs in the queue. To do this, Rega coordinates with the HPC's inbuilt Slurm scheduler to submit jobs to the queue. Since Rega has the ability to calculate many sites at once, the most



efficient manner to submit these jobs to the HPC is via a Slurm array, where a script is submitted that uses a list of names and locations of each calculation and the resources required. A Slurm array script is generated for each batch of different computational resource requirements and uploaded to the HPC. Rega submits each array, and the jobs are scheduled for calculation.

In order to monitor the progress of the calculation, the directory of the calculation must be paired with the JobID assigned by Slurm. As part of the array script, a Slurm file is generated in the remote home directory containing the path of the calculation associated with that JobID. Rega goes through these files and maps the file path with the JobID and notes the calculation status in the Slurm queue as either pending, running or not present (therefore completed). Rega periodically checks the Slurm queue for the status of each job for overall calculation status but also monitors the progress of the transition state search. Firstly, the output file is downloaded periodically and parsed to find the most recent output geometry through the course of the search. This geometry undergoes similar distance checks as before in the semi-empirical calculations and determines whether the calculation is still on course for finding the HF transition state or if it is converging to a minimum. In the latter case, the type of minimum is noted (i.e. if the C-C bond of the added radical is elongating then the starting geometry had too great of a reactant geometry character and if the C-C bond is too short then there is too great of a product character in the input geometry). The job is killed in the Slurm queue, and a modified input geometry is generated in response to the nature of the minima convergence seen in the first transition state search attempt.

This monitoring of calculation progress and early termination of failing transition state searches is done to aid calculation efficiency of Rega, since the inbuilt optimisation algorithms typically take many SCF cycles before crashing, noting the point of failure being convergence to a minimum rather than a maximum. The SCF cycles taken for the inbuilt algorithm to crash are better spent working on convergence to the true transition state and so these calculations are manually killed and restarted by Rega, saving time.

#### 4.2.4 Resubmission and Data Collation

Once calculations are completed the frequency output of the final geometry is parsed. If a true HF transition state is found, the output file is downloaded and the site is noted as completed in Rega. If there are additional imaginary frequencies in the first TS state search, the final geometry from this search is resubmitted via a new Slurm array for a second TS calculation with tightened convergence criteria. This resubmission is conducted only after each calculation in the queue is checked to reduce the number of Slurm arrays being submitted, with each array being binned for the compute requirements as before. With calculations that were killed due to convergence to minima, these “first” transition state search attempts with tweaked input geometry are submitted with any calculations that require a second search via the same Slurm array. These steps are taken to reduce the manual intervention typically required for these calculations and remove the need for detailed knowledge of the QM chemistry software, making the generation of machine learning datasets

more accessible to the average chemist. In the rare instance that a calculation does not converge to a successful transition state after the second, tighter criteria search, the site is assigned a “No TS found” tag and Rega moves onto the next compound. This is due to the thought that the system resources utilised for calculation on this site are better served on performing a TS search on a new compound rather than using greater computational expense to find this single failed transition state.

Once all calculations are completed, a store data function collates all compounds and sites of reaction and gathers the transition state energy, reactant energy and property calculations and stores them in a nested dictionary. This is output into a .json file for the entire batch of compounds and a .csv file is written for each compound dictionary. Rega is also robust to the calculation of the same substrate with a different carbon radical, as the previous compound dictionary is appended with the newly calculated energies.

## 4.3 Graphical User Interface and Database Management

To improve the ease of use of Rega, a graphical user interface (GUI) was built. This enables users with no experience in computer science and command line interface navigation to run calculations, increasing the likelihood of wide adoption in an industrial setting. An overview of the graphical user interface is given in **Figure 4.4**. Firstly, the user is prompted whether they would like to calculate the regioselectivity of a single compound or a list of multiple compounds. If calculating a single

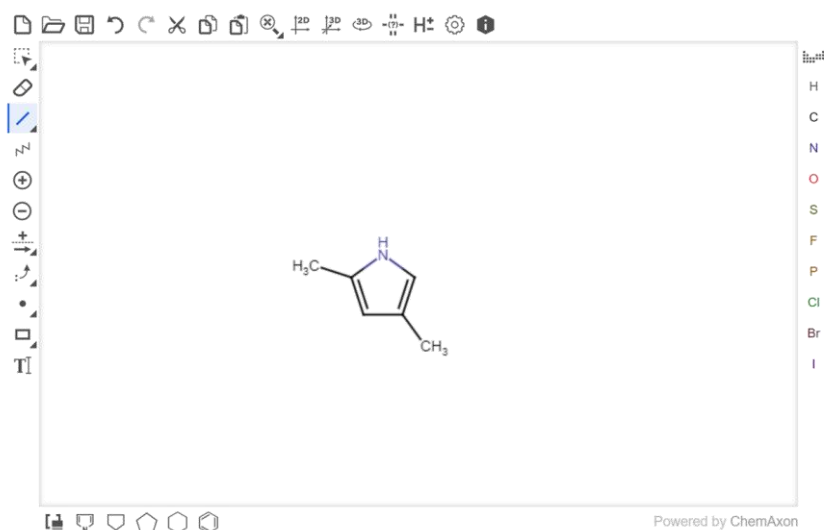
compound, the user is provided with a molecular editor in which they can draw the compound of interest directly. Once drawn, the SMILES string of the compound is generated and the SMILES box in the form below is populated. The user then gives the name of the compound and selects which radical they would like to study from a drop-down list of  $\cdot\text{CF}_3$ ,  $\cdot\text{CF}_2\text{H}$  and  $\cdot\text{iPr}$ . Lastly, the user selects whether they would like to calculate energies using the more expensive but accurate HPC calculation functions. Once submitted, Rega is started with all required information given.

If the user requires a batch of multiple substrates to be calculated, a .csv file is uploaded to the GUI with instructions provided on screen on the format of the file required for the successful submission of these compounds. Once calculations are complete, the compound is displayed alongside a table showing the activation energies for each labelled site of reaction and the corresponding regioisomeric ratio (**Figure 4.4**)

In order to further increase calculation efficiency, the database of previously calculated compounds is stored locally and allows for easy retrieval of data. If the user submits a compound that has already been through the Rega program, its activation energy and regioisomeric ratio are gathered from the database and displayed to the user instantly, avoiding unnecessary compute time. Also, if the same substrate is calculated with a different radical, the database is updated with this new information. Users can then download the results to a .json file for easy post-processing and incorporation into machine learning datasets.

## Single Compound

For single compound calculation either draw a molecule or input the SMILES below



Transfer/Update SMILES

Name of Compound: 2,4-Dimethylpyrrole

SMILES string of Compound: CC1=CC(C)=CN1

Radical to add (cf3, cf2h, ipr): cf3

Calculate accurate energies using the HPC?: ☐

Submit Single

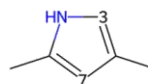
Rega

Calculations Complete

Calculate Another Compound

Compound: 2,4-Dimethylpyrrole  
Radical: cf3

Site	Activation Energy (kcal.mol <sup>-1</sup> )	Regioisomeric Ratio
3	3.4	280.8
7	7.3	1.0



Download Results

Figure 4.4 Graphical user interface for single compound calculation on Rega.

## 4.4 Conclusion

Rega offers a new methodology for the high throughput calculation of large numbers of compounds in a robust manner with minimal manual intervention. Rega is designed to be easy to use for the non-computational chemist and gives the user the ability to locate transition states for a wide variety of substrates in difficult-to-model reaction systems with a fraction of the time and resources typically required. The template-based workflow allows easy adaptation to different reaction systems and simple toggle switches within Rega enable the user to easily specify the computational software and HPC service of choice. Rega's ability to locate previously calculated compounds within the database enables instant readout of results where the same substrate/radical pair has been generated previously. When the same substrate is being calculated with a different radical, the database updates the information for that compound ready for easy retrieval upon request.

The Rega calculation framework enables the generation of a large dataset of calculated regioselectivities for the generation of machine learning models. These results are presented in the next chapter.

# Chapter 5 - Dataset Generation & Machine Learning

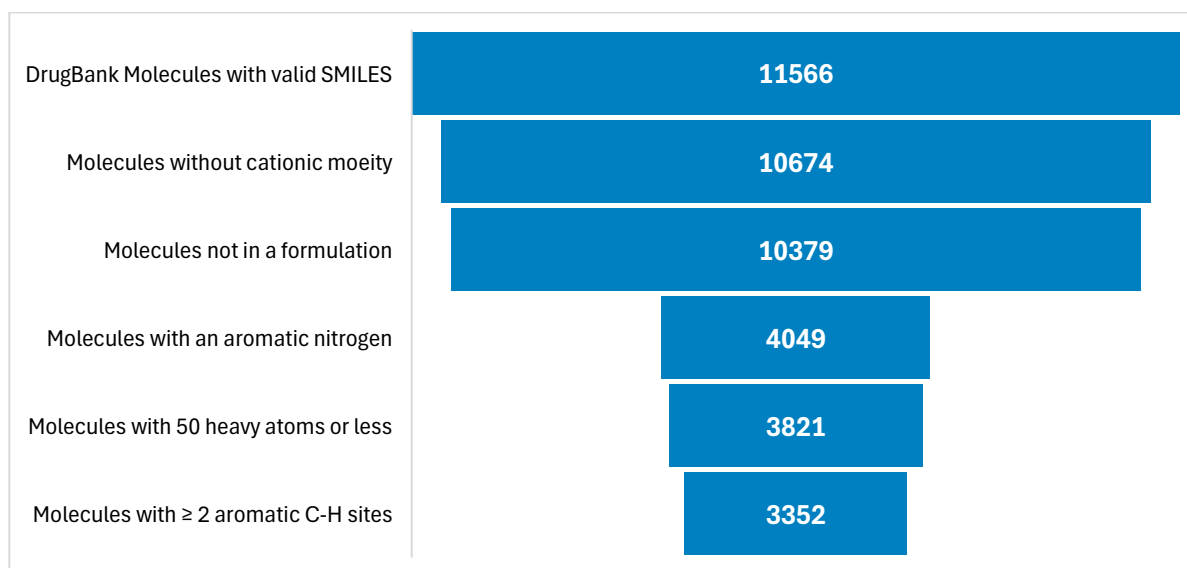
## 5.1 Dataset Generation

In this section of the chapter, we discuss the steps taken to generate an appropriate artificial dataset for the zinc-sulfinate mediated C-H functionalisation reaction.

### 5.1.1 Dataset Curation

To generate the necessary dataset for the C-H functionalisation reaction a starting database containing a diverse array of drug-like compounds was selected to maximise the domain of applicability of the models. The DrugBank database was used,<sup>226</sup> which contains over 16,500 compounds including 2752 approved small molecule drugs, 6723 experimental drugs and 1600 biologics. This dataset was sampled ensuring a broad and uniform coverage of chemical space to ensure consistent model performance irrespective of training/test set split. From the original 16,500 compounds, several filters were applied to make the data more representative of the classes of structures where this reaction may be considered. Firstly, permanently charged species were removed, due to the difficulty in prediction of protonated species; the position of protonation is challenging to

predict, introducing additional sources of error. Thus, cationic species are beyond the scope of our investigation. Also, compounds that appeared as a formulation of multiple molecules were removed, since the focus is on the regiochemistry of the drug compound and not interactions with other compounds in the formulation. The number of heavy atoms in each compound was restricted to 50, recognising that drug-like molecules typically have a molecular weight of less than 500.<sup>227</sup> Also, since this reaction typically acts on nitrogen-containing heteroaromatic rings, any compounds with no aromatic nitrogen atoms were removed. Lastly, any compounds with less than two potential sites of reaction (in this case, aromatic C–H carbon atoms) were removed, as only one potential site of reaction would lead to only a single regioisomer. **Figure 5.1** shows the number of remaining compounds from the DrugBank database remaining after each of the filtering criteria have been met.



**Figure 5.1** Funnel plot of the number of compounds remaining after each filtering step.



After this filtering, the remaining 3352 compounds were clustered using the Tanimoto coefficient<sup>228</sup> with a similarity threshold of 0.7, to generate a large number of small clusters. These clusters were sampled randomly, to ensure the final dataset contains compounds from all areas of chemical space contained in the original set. A set of over 500 compounds were selected, containing more than 2800 potential sites of reaction. To measure the diversity of the dataset and therefore the potential applicability of this set for use in regioselectivity prediction in drug discovery, the pairwise diversity was calculated using the Tanimoto coefficient. The calculated average pairwise coefficient of 0.59 shows a high diversity across the drug-like space. Therefore, modelling the regioselective behaviour of this reaction on this set would have a large domain of applicability.

### 5.1.2 Large-Scale Calculations and Data Preprocessing

Once the dataset was collated, AM1 transition states were located with Rega and HPC input files were generated. The HPC facility chosen was Sulis, a Tier 2 cluster computing service located at the University of Warwick. This system is comprised of 28,244 high-performance CPU cores with 4 GB of memory allocated per core, making it ideally suited to running the large number of transition state calculations required to generate this dataset. To circumvent the stricter multi-factor authentication criteria employed by the Sulis HPC facility, a pared-down version of Rega was used that terminates after the upload and submission of each calculation. This was due to the inability of Rega to continue establishing connections with Sulis overnight,

preventing the full automation ability of Rega previously used other HPC facilities. Sulis was chosen due to the greater computational resources available, and the increased SCF cycle speed seen when benchmarking facilities on identical systems.

Once finished, calculation output files were downloaded using a modified version of Rega and then checked for successful completion. When training the model, it is imperative that no spurious data points are included in training to maximise the probability of finding the correct relationship between site environment and activation energy. Any activation energies that gave either a negative value or a large value (greater than 100 kcal mol<sup>-1</sup>) were removed along with any other sites that may have a seemingly appropriate activation energy, as this was seen to be a “bad compound” removing the possibility of poor data being used in training. Any compounds where a single site was not able to be successfully calculated were also removed as this was also considered to be a compound needing further attention. After this final filtering process, the dataset to be used for modelling comprised 490 compounds containing 2744 potential sites of reaction. If the full version of Rega was allowed access to Sulis, the number of completed compounds would have been greater, as its ability to resubmit failing calculations and resubmission of sites that finish the first TS search close to the true transition state would greatly increase capture rate. However, Sulis access permissions/time pressure would not allow for this full implementation.

Once collated, data was split in an 80:10:10 ratio of training: validation: test sets in five separate folds for cross-validation. Data were split by compound rather than by

site, meaning all sites within the compound are contained in an individual set and no sites are split across training and test, ensuring no data leakage and therefore avoiding artificial boosting of model performance.

### 5.1.3 Traditional Modelling

Initial modelling efforts were based on a traditional QSAR regression model approach, using descriptors generated by RDKit for each trifluoromethylated product SMILES structure, since this structure is the only unique element for each row in the set. Using Scikit-learn,<sup>229</sup> models were built using principal component analysis to identify key descriptors in the relationship between structure and calculated activation energy. The primary metric used to assess model performance is the root mean squared error (RMSE). Since in experiment, we see multiple regioisomers formed in some compounds, it is clear that in order for a model to be deemed successful, it must be able to detect the subtle differences in activation energy between sites in a compound, some of which have an activation energy difference of 0.1 kcal mol<sup>-1</sup>. Therefore, an RMSE of below 1 kcal mol<sup>-1</sup> is required. Another measure of performance is the coefficient of determination (also known as  $R^2$ ). This measures the proportion of variation in the dependent variable that can be attributed to the independent variable and values range between zero and one; where 1 represents a perfect correlation and 0 indicates no dependency between variables.

**Table 5.1** shows the results of typical QSAR modelling efforts on the artificial activation energy dataset. Random forests gave the greatest model performance

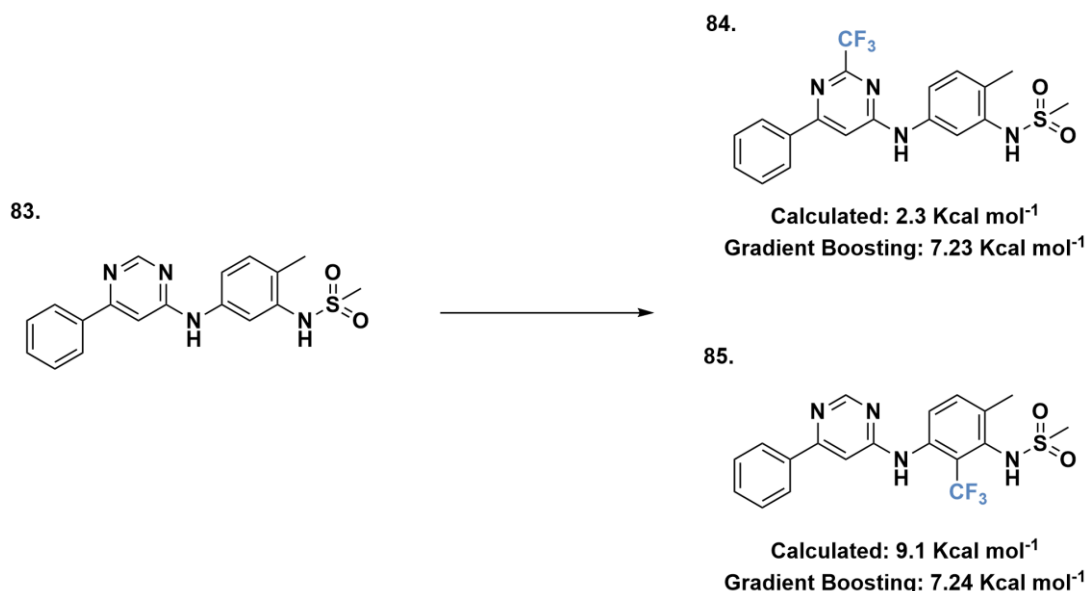
with an average RMSE of 3.04 kcal mol<sup>-1</sup>,  $R^2$  of 0.06 and variance of 9.3. Individual set performance was consistent across the five folds and set four and five RMSE was below 3 kcal mol<sup>-1</sup>, the only examples seen in these modelling efforts. The poorest performing method was Gaussian Processes with the constant kernel where we see an average RMSE of over 10,000 kcal mol<sup>-1</sup>, due to the poor result seen in set 4. Excluding this result the RMSE exceeded 44 kcal mol<sup>-1</sup>, making it unsuitable for accurate predictions of regioselectivity.

Set	Linear Regression	Random Forests	XGBoost	GP_White Kernel	GP_Matern	GP_Rational Quadratic	GP_Constant Kernel	GP_RBF
1	3.15 (0.06)	3.01 (0.10)	3.20 (0.07)	3.15 (0.06)	7.52 (0.001)	3.05 (0.08)	135.00 (0.01)	7.62 (0.0005)
2	3.59 (0.04)	3.22 (0.06)	3.38 (0.04)	3.50 (0.04)	4.41 (0.001)	3.30 (0.02)	14.36 (0.01)	6.77 (0.02)
3	3.04 (0.08)	3.03 (0.06)	3.37 (0.01)	3.07 (0.05)	5.53 (0.003)	3.03 (0.04)	12.39 (0.01)	7.87 (0.01)
4	4.01 (0.0005)	2.98 (0.01)	3.19 (0.03)	6113.44 (0.08)	3.54 (0.0001)	3.09 (0.04)	>10000 (0.08)	4.25 (0.005)
5	3.58 (0.001)	2.97 (0.02)	3.20 (0.01)	3.38 (0.003)	5.30 (0.02)	2.94 (0.03)	15.07 (0.02)	7.70 (N/A)
Average	<b>3.48</b> <b>(0.02)</b> <b>{12.3}</b>	<b>3.04</b> <b>(0.06)</b> <b>{9.3}</b>	<b>3.27</b> <b>(0.03)</b> <b>{10.7}</b>	<b>1225.31</b> <b>(0.01)</b> <b>{74143}</b>	<b>5.26</b> <b>(0.002)</b> <b>{17.6}</b>	<b>3.08</b> <b>(0.04)</b> <b>{9.5}</b>	<b>&gt;10000</b> <b>(0.01)</b> <b>{7.6 E+18}</b>	<b>6.84</b> <b>(0.002)</b> <b>{16.5}</b>

**Table 5.1:** RMSE ( $R^2$ ) {Variance} of traditional QSAR models across five-fold cross-validation.

Looking at the predictive behaviour of each modelling method on the test set, there appears to be no consistent pattern in the site environment assigned the lowest

activation energy. **Figure 5.2** shows that calculated activation energies agree with regioselective behaviour seen in other literature examples, where sites proximal to heteroaromatic nitrogens are favoured. XGBoost does not detect this behaviour and instead assigns two sites with very different calculated labilities the same activation energy. This shows that the characteristics of the literature and calculated preferred site of reaction have not been captured in modelling. This example is consistent across different traditional modelling methods and may be due to a number of factors. Firstly, the features generated from RDKit may not be sufficient to capture the differences in site environment correctly. Secondly, while the dataset used for calculation is more broadly described as the drug-like chemical space, it may be that the variety of compounds in this drug space is too diverse for the model to effectively capture the steric and electronic properties of each potential site of reaction that give rise to the regioselective behaviours seen in calculations.



**Figure 5.2** Example prediction vs calculation for gradient boosting (XGBoost) method.

## 5.2 ChemProp Modelling

Following the poor performance of previous modelling efforts, graph neural networks were employed to understand if the graphical representation of each structure would enable better chemical intuition and therefore gain a greater understanding of the site environment's reactivity.

### 5.2.1 Chemprop Regression Models

For ChemProp modelling, each activation energy is paired with the associated reaction taking place, with the substrate and product SMILES joined by the ">>" symbol. This "reaction SMILES" representation attempts to assign the activation energy for the direct chemical transformation taking place at each site and graphs the chemical connections/disconnections taking place. The data were again split molecule by molecule to avoid any data leakage of some sites of a particular molecule appearing in the training set and other sites of the same molecule appearing in the test set. Data were split in an 80:10:10 ratio of training:validation:test sets corresponding to an average number of 2165 sites in the training set, 273 in the validation set and 268 sites in the test set. Datasets were loaded into ChemProp and additional RDKit features were generated to complement the graph representation.

Set	Reaction SMILES	Mapped Reaction SMILES	Mapped + Morgan Fingerprints
1	2.97	2.86	2.98
2	3.01	3.17	3.22
3	3.08	3.01	3.07
4	3.05	2.94	2.92
5	2.88	2.87	2.84
Average	2.99 (0.086)	2.97 (0.029)	3.00 (0.011)

**Table 5.2** ChemProp Regression RMSE results for both typical reaction SMILES and atom-mapped reaction SMILES.

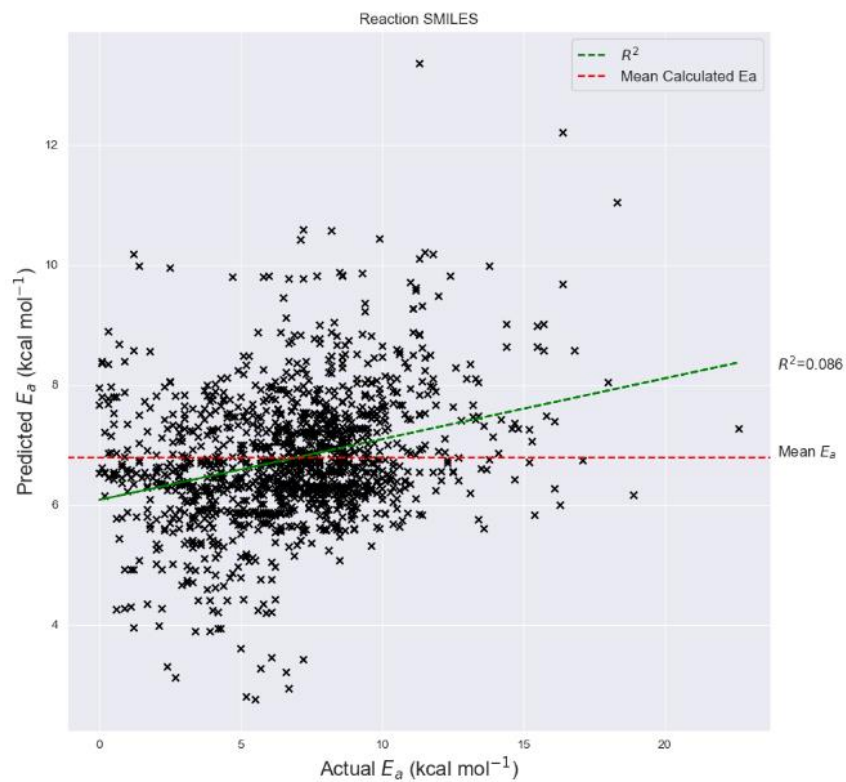
The results of the first regression modelling attempt are shown in **Table 5.2**. While we do see an improvement in RMSE over other previous modelling attempts, falling below 3 kcal mol<sup>-1</sup>, we do not see a substantial enough increase in performance to consider the prediction outputs reliable for deployment. Another modification attempted on the dataset was the utilisation of atom-mapped reaction SMILES. These SMILES strings contain atom number assignments within the string of both reactant and product structures. The aim was to investigate whether providing more explicit information on the atom number of the substrate taking part in the reaction would help boost model performance. **Table 5.2** shows that while we did see an improvement in RMSE for cross-validation sets 1,3,4 and 5, these gains were not substantial enough to warrant the model effective in providing an accurate prediction of regioselectivity. This is illustrated further in Figure **5.3a** where we see an  $R^2$  correlation of 0.086 for the typical reaction SMILES representation and 0.029

for atom-mapped reaction SMILES. The addition of Morgan fingerprints<sup>230</sup> in the feature set improved sets 4 and 5 RMSE but sets 1 and 2 performances increased the average RMSE over other methods attempted.

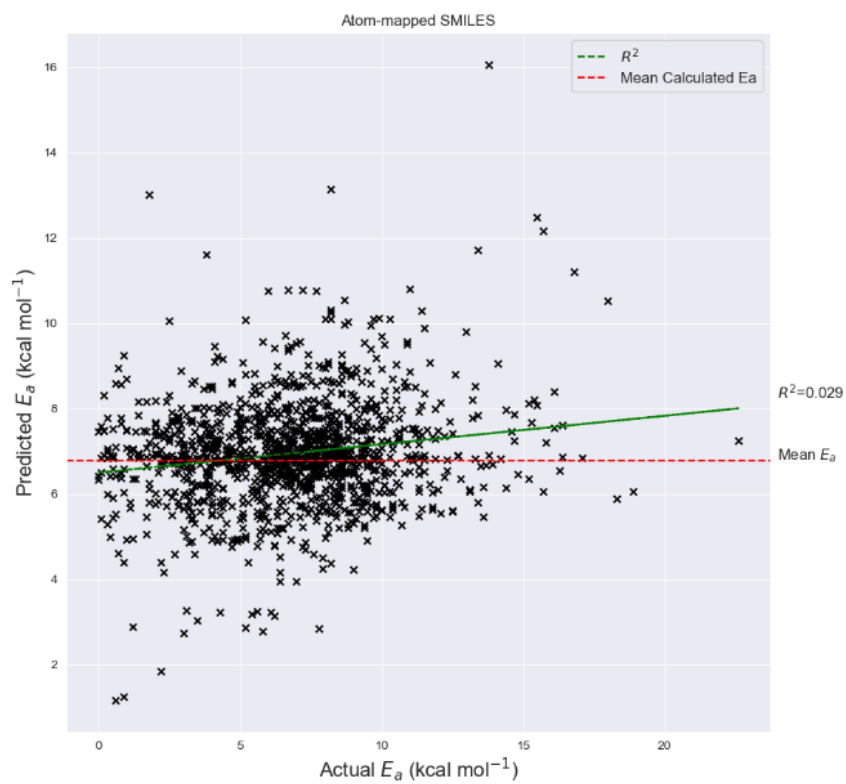
Other approaches were attempted to boost model performance. One such attempt was the splitting of the dataset randomly rather than by compound to see if there was any improvement, which was unfortunately also unsuccessful. Another method attempted was the removal of the GNN component of ChemProp and the direct feeding of reactant and product features into the feed-forward network (FFN) layer. This modification did not yield any improvement in performance above the baseline models discussed earlier. Lastly, the addition of a range of QM descriptors such as Fukui indices and Mulliken charges as well as site-specific electron density and electrostatic potential saw slight improvements in model performance across certain cross-validation splits but reduction in others. This led to no overall improvement in model performance over the classic implementation of ChemProp previously attempted.



a



b)



**Figure 5.3:** Calculated vs predicted activation energies for the ChemProp regression model using typical reaction SMILES.

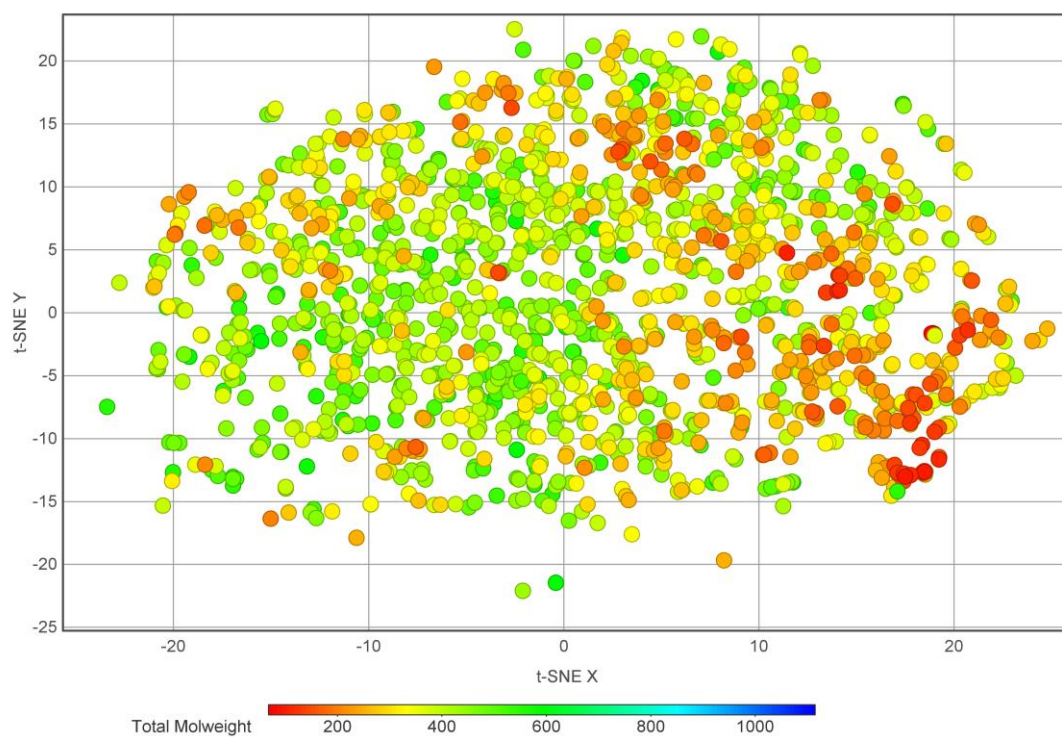
### 5.2.2 ChemProp Classification Models

Since the task of regression on this data appeared to be too difficult to make accurate predictions of activation energy, the task was converted to a classification problem. This new dataset representation used a binary classification of whether the site underwent functionalisation. For each compound, the lowest activation energy site was assigned a 1 and the other sites of reaction were assigned a 0. Each input row is the compound in question and the predicted output is an array of length  $n$ , where  $n$  represents the number of potential sites of reaction, with 1 assigned to the most likely site of reaction and all other array entries assigned 0. Top-k accuracy (a metric for scoring model accuracy based on correct classification after  $k$  predictions) was measured for the binary classification model (**Table 5.3**). Top-1 and top-2 accuracies were only 31.8% and 59.3% respectively, not accurate enough to be considered a viable model.

Top -k	Accuracy (%)
1	31.8
2	59.3
3	71.2
4	78.4
5	79.7

**Table 5.3:** Top-K accuracy for the binary classification ChemProp model.

ChemProp modelling has shown that even the most viable models do not meet the accuracy requirements for deployment as a useful model for predicting regioselectivity. It appears that the dataset generated from Drugbank was too diverse to sufficiently learn the relationship between site environment and site lability for this C-H functionalisation reaction. While the compounds included in the dataset are all from the 'drug-like' chemical space, it seems the sampling of only a few compounds from each cluster was not enough to fully capture the features of these similar structures and the effects of these on activation energy. Also, the similarity between the structures in different clusters was overestimated, and so any learned relationship between structure and site lability for a given structure could not be effectively extrapolated to compounds in other clusters. This means that while there may be a small signal obtained from compounds within a single cluster, the amount of information gathered was not sufficient to understand the real relationship between a site's steric/electronic environment and its lability in this reaction for more diverse species in different clusters. **Figure 5.4** shows the wide array of chemical space covered by the DrugBank database. We do not see any distinct grouping of compounds in any given region of chemical space, making effective sampling from this dataset challenging. Due to this diversity, in future, sampling should be carried out on a smaller region of this chemical space in larger quantities in order to gain a greater understanding of the relationship between those structures and regioselective behaviour before expanding into new regions of chemical space.



**Figure 5.4** t-SNE plot of the clustered DrugBank dataset.

## Chapter 6 - Conclusions & Outlook

In conclusion, the prediction of C-H functionalisation regiochemistry is a challenge preventing deployment in industrial applications, particularly those reactions involving complex reaction pathways with reactive intermediates. While there have been many different approaches to predicting these reactions' regioselectivity, both data scarcity and reaction condition variation make the understanding of a substrate's properties that contribute to its regioselective behaviour challenging. Two distinctly different methods were trialled in the prediction of the sulfinate-mediated class of C-H functionalisation reactions in this work, the electronic population on each potential site of reaction of the ground state compound, and the nature of each transition state. While preliminary investigations saw promise in the atom-centred charges methodology, its application to larger more drug-like structures indicative of those typically used in industry showed its shortcomings. This was due to the lack of any steric contribution to regioselectivity in this method which plays a key role in the determination of site lability. Activation energy on the other hand showed great promise in the prediction of regiochemistry for both simple substituted heteroarenes and more complex substrates containing multiple heteroaromatic moieties. This approach accounts for both electronic and steric factors that affect a site's susceptibility to radical attack and maps the reaction pathway, allowing for accurate determination of these kinetically controlled reactions. Since this methodology poses a challenge of significant computational

expense, QM calculation screening was conducted to evaluate the method with the greatest balance of accuracy at the lowest cost. The result of this investigation led to the use of the HF/6-31G\* method as the theory of choice which showed great regioselectivity predictive performance in a fraction of the time of the more expensive DFT methods.

Once HF/6-31G\* was validated, it was decided that in an effort to avoid the computational cost for the end-user, machine learning should be employed to predict this reaction's regioselectivity through the prediction of a site's activation energy. To do this, an artificial dataset was curated since the volume of high-quality experimental data was insufficient for any machine learning technique to accurately predict these energies. To increase computational efficiency further, an automated high-throughput transition state location program known as Rega was developed to generate this large dataset of calculated regioselectivities for a wide array of different drug-like compounds. This Rega workflow is unique in its ability to schedule and monitor hundreds of calculations simultaneously on remote HPC clusters and correct any transition state searches that are converging to a minimum on the PES or have not found the true saddle point. The modularity of the Rega program allows for simple adaptation to different HPC clusters and new reaction schemes, enabling broader applications to new areas of chemistry where the prediction of regioselectivity is a challenge. The easy-to-use graphical user interface further removes the barrier to entry for this program and enables laboratory chemists with little to no knowledge of computational chemistry to make these regioselectivity predictions by drawing their compound of interest and allowing Rega to perform

these complex calculations in the background. The user's ability to save these calculations in an easily digestible format allows for easy incorporation into machine learning datasets and the constantly updating database of successful calculations enables instant retrieval of data for any compounds previously run through Rega, saving time.

With this dataset of over 480 complex drug-like compounds successfully calculated, it was then used to build both classical machine learning models and graph neural networks to attempt to predict this regioselectivity in both regression and classification tasks. While model performance was not adequate to release a suitably useable model, the Rega program allows for further work to be done in future to solve this problem.

It was proposed that should further work be done to model activation energy for this C-H functionalisation reaction, more data should be obtained from a smaller region of chemical space. When performing clustering using the Tanimoto coefficient of molecular similarity as described in 5.1.1, instead of sampling a small collection of a few compounds from many clusters in the drug-like space, all structures from a single cluster should be calculated and a model trained on this subsection of chemical space. Doing so would greatly increase the understanding of the structure/function relationship between potential sites of reaction in this region of chemical space and maximise the possibility of generating an effective activation energy prediction model. From that new model, additional data on full clusters of compounds can be added using Rega and the model retrained, giving a higher chance of the signal

obtained in the training on the first cluster to be applicable to the next cluster of compounds. This chemical space expansion approach yields the best chance of producing an effective model to accurately predict regioselectivity in the vast drug-like space.

The potential to incorporate active learning into the model is another exciting avenue for exploration. In this work, a finalised model would include an uncertainty metric in the output of its prediction. From this, if a compound has a low certainty of correctly predicting the regioselectivity (if the compound is too dissimilar to the compounds the model has been trained upon), then the user would be prompted to calculate the activation energy for that specific input compound using the Rega automated transition state search program. With that new data point generated, the model would then retrain, gaining new information about regioselective patterns in that new region of chemical space. In short, the model would improve over time as more people used the tool.

Another potential direction of research is the deployment of the Rega platform on new reaction systems. Given the modular nature of the program, this facilitates the modification of transition state templates and checking criteria to different reaction schemes to tackle new regioselectivity problems in organic chemistry. It is hoped that in future this work will be continued to advance the field of machine learning-derived regioselectivity prediction further.



# Chapter 7 - Appendix

## 7.1 General Experimental Information

All chemicals were purchased from Sigma Aldrich, Alfa Aesar, Fisher Scientific, Fluorochem or Manchester Organics.

Analytical thin-layer chromatography was carried out on glass or aluminium-backed plates coated with Merck Kieselgel 60 GF254 purchased from Merck.

**NMR** spectra were recorded at 298 K using Bruker AV(III)400, AV400 (400 MHz  $^1\text{H}$  frequency, 100 MHz  $^{13}\text{C}$  frequency) or Bruker AV(III)500 (AV400 (400 MHz  $^1\text{H}$  frequency, 100 MHz  $^{13}\text{C}$  frequency, equipped with a cryoprobe). Chemical shifts are quoted in parts per million (ppm), referenced to residual chloroform (7.26 ppm for  $^1\text{H}$  NMR, 77.16 ppm for  $^{13}\text{C}$  NMR), dimethylsulfoxide (2.50 ppm for  $^1\text{H}$  NMR, 39.51 ppm for  $^{13}\text{C}$  NMR), and methanol (3.31 ppm for  $^1\text{H}$  NMR, 49.00 ppm for  $^{13}\text{C}$  NMR) as internal standards and coupling constants,  $J$ , are quoted in Hz. Multiplicities are as follows: s – singlet, br s – broad singlet, m – multiplet, d – doublet, dd – doublet of doublets, ddd – doublet of doublet of doublets, dt – doublet of triplets, t – triplet, q – quartet.

**Under reduced pressure** refers to the use of a Vaccubrand CVC 3000 vacuum pump to remove solvent under reduced pressure on a Büchi Rotavapor R-3000 or Heidolph Vei-Vap Value G3 apparatus, with a water bath at 40 °C.

**TLC** plates were visualised under UV light (254 or 365 nm) and/ or stained with the appropriate staining solution. The staining solution is reported when used: either basic aqueous potassium permanganate or ethanolic cerium phosphomolybdate.

**Column chromatography** was carried out using Interchim Puriflash pre-packed silica gel columns, eluting with the aid of an Asynt chromatography pump or Biotage SP1 chromatography system.

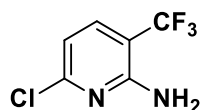
**Melting points** were measured on a Stuart SMP20 digital melting point apparatus and are reported to the nearest degree, uncorrected.

**Mass Spectrometric** analyses at the School of Chemistry, GlaxoSmithKline Carbon Neutral Laboratories, University of Nottingham were recorded on a Bruker MicroTOF 61 mass spectrometer using electrospray ionization (ESI). *m/z* values are reported in Daltons. For GCMS analysis the JEOL AccuTOF GCX mass spectrometer was used using electron ionisation.

**Infrared spectra** were recorded using a Bruker Alpha Platinum ATR single reflection diamond module spectrometer over the range of 4000 – 600 cm<sup>-1</sup>.

## 7.2 Synthesis of Trifluoromethylated Literature Compounds

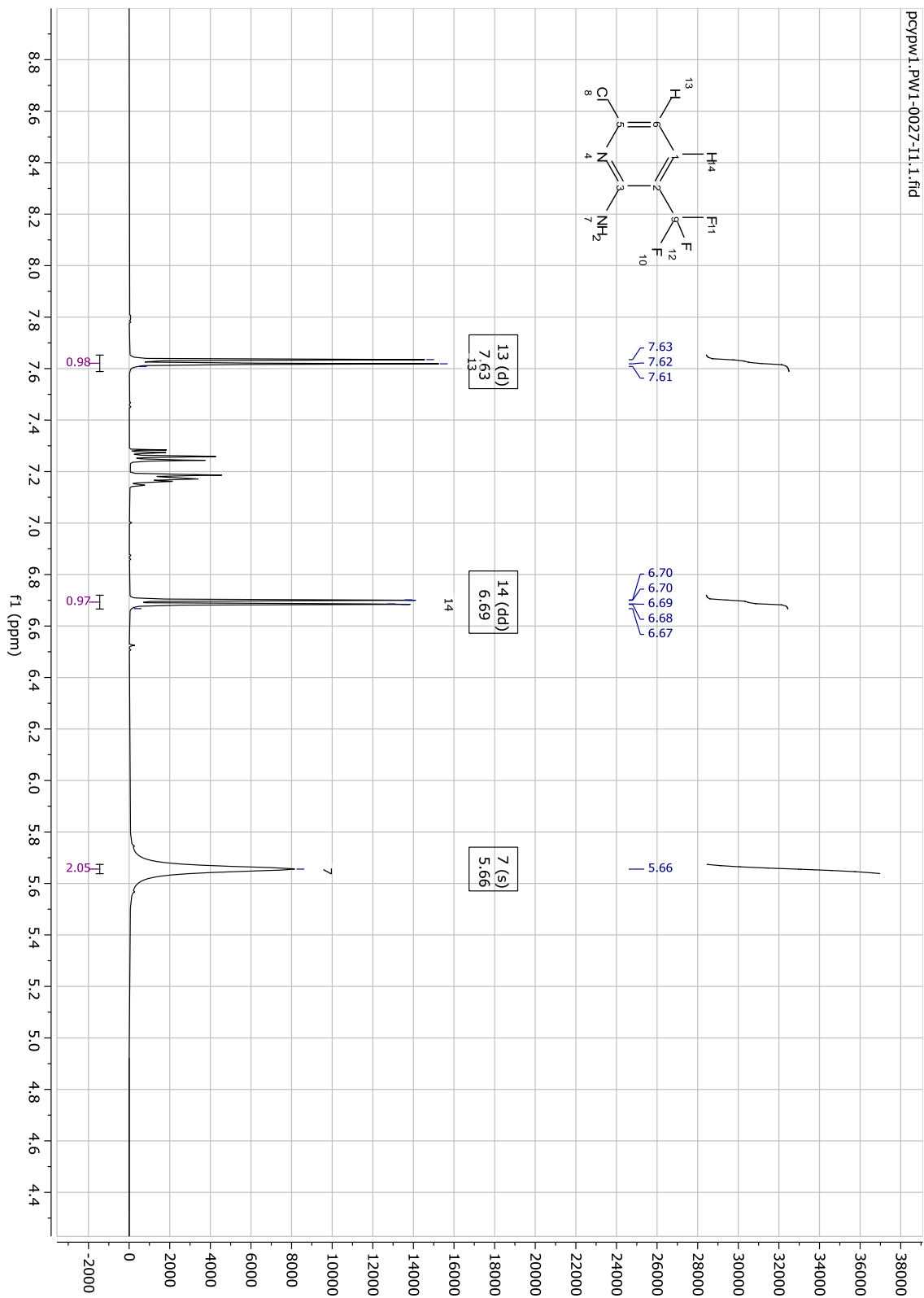
### 6-chloro-3-(trifluoromethyl)pyridin-2-amine

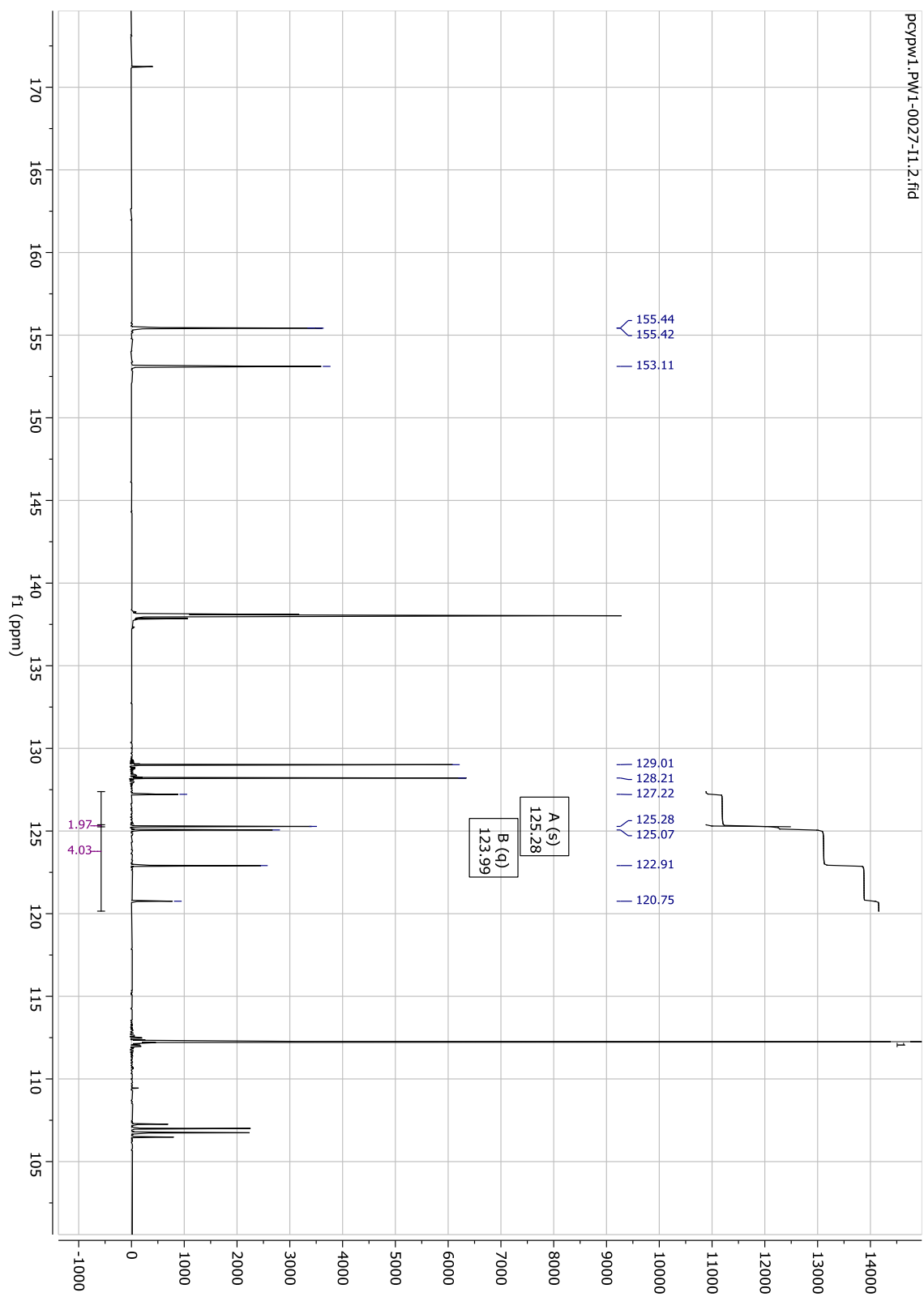


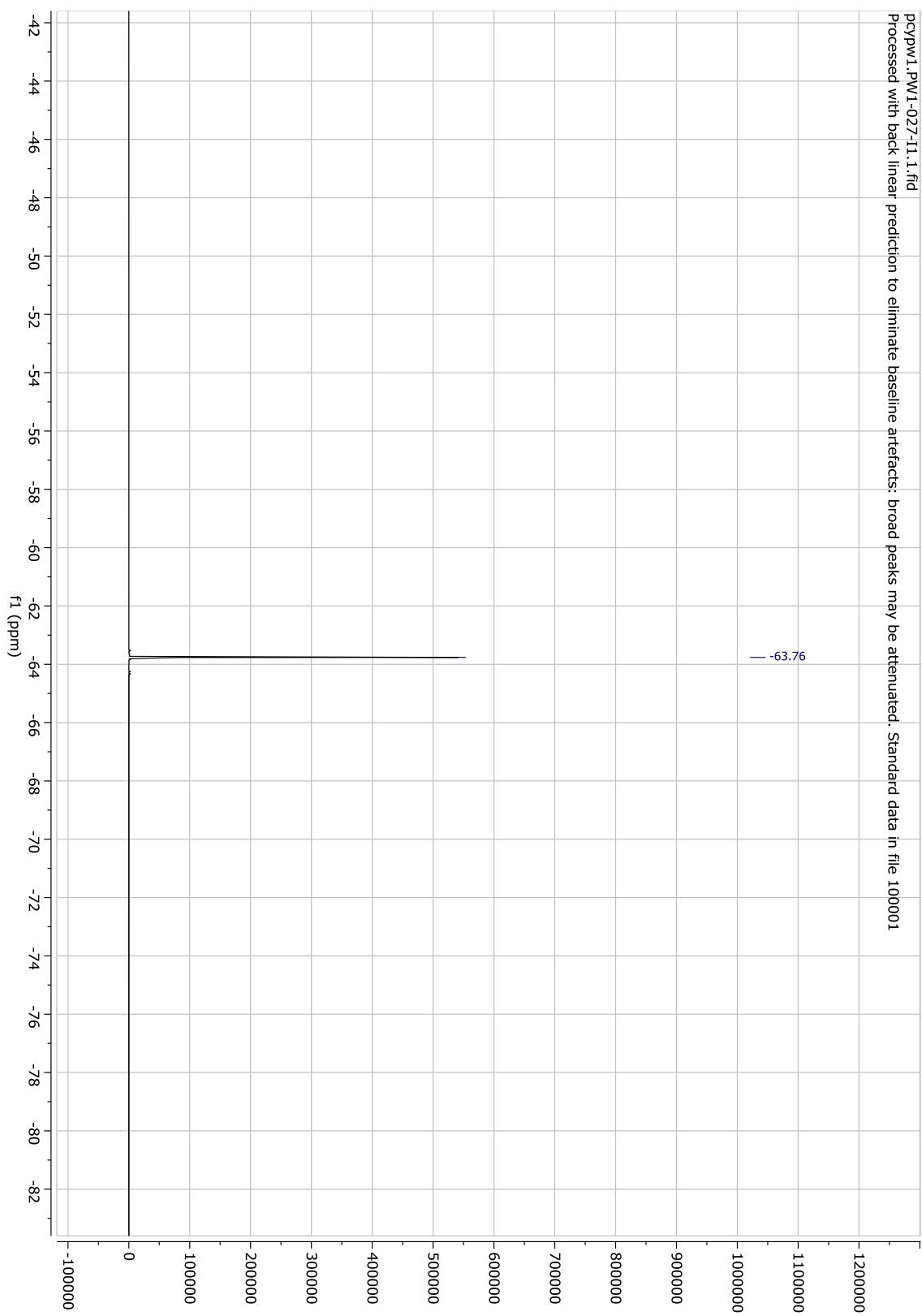
86

To a mixture of dichloromethane (35 mL) and water (15 mL) was added zinc trifluoromethanesulfinate (2.6 g, 7.8 mmol) and 6-chloropyridin-2-amine (0.5 g, 3.9 mmol). The mixture was cooled and *tert*-Butyl hydroperoxide (1.3 mL, 11.7 mmol) was added dropwise over the course of 5 minutes and the reaction mixture was heated at 50 °C for 48 hours. After this time, the reaction mixture was quenched with EDTA:sodium hydrogen carbonate (1:1 mixture of a 4 M aqueous solution and a saturated aqueous solution) (50 mL) and the organic layer dried over MgSO<sub>4</sub>, filtered and evaporated under reduced pressure. The residue was diluted with dichloromethane and adsorbed onto silica gel. Purification by silica gel chromatography, eluting with ethyl acetate, toluene and pentane (4.85:0.15:95), provided the *title compound* (226 mg, 31%) as a yellow solid: **mp.** 109-110 °C; **R<sub>f</sub>** 0.43 (10% EtOAc:90% Pentane); **FT-ATR**  $\nu_{\text{max}}$  3519, 3307, 3179, 1663, 1592, 1567, 1462, 1310, 1267, 1209, 1157, 1093, 1065, 1016, 960, 932, 806, 769, 757; **<sup>1</sup>H NMR** (400 MHz, CDCl<sub>3</sub>)  $\delta$  7.66 (d, *J* = 8.0 Hz, 1H), 6.75 (dd, *J* = 8.0, 0.9 Hz, 1H), 5.18 (s, 2H); **<sup>13</sup>C NMR** (126 MHz, CDCl<sub>3</sub>)  $\delta$  155.4, 153.1, 129.0, 128.2, 125.3, 124.0 (q, *J* = 271.2 Hz); **<sup>19</sup>F**

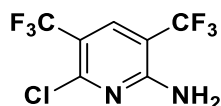
**NMR (376 MHz, CDCl<sub>3</sub>)**  $\delta$  -63.61; **MS**  $m/z$  (EI) calcd for C<sub>6</sub>H<sub>4</sub>ClF<sub>3</sub>N<sub>2</sub> [M<sup>+</sup>] requires 196.0015, found 196.0004.







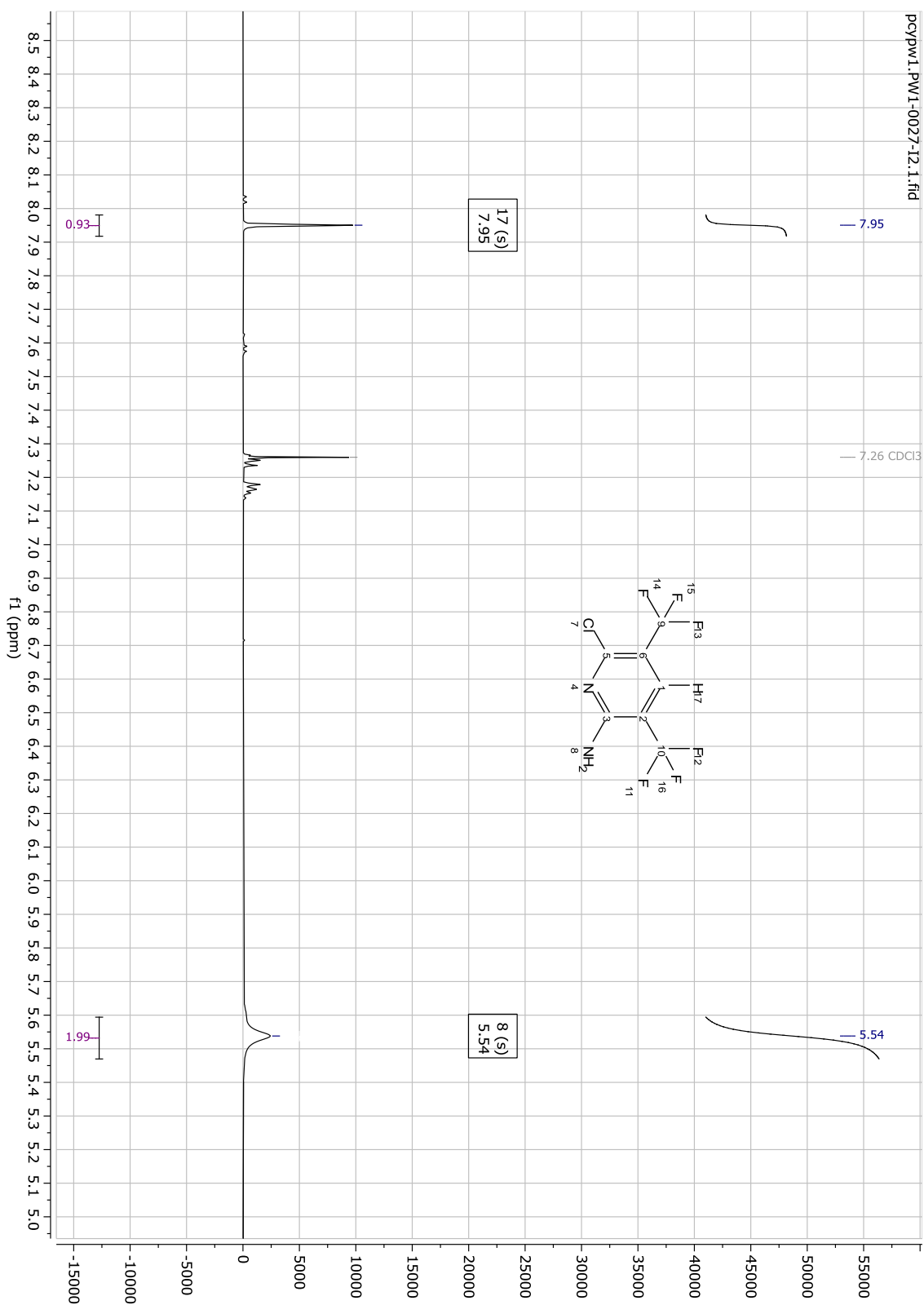
**6-chloro-3,5-bis(trifluoromethyl)pyridin-2-amine**

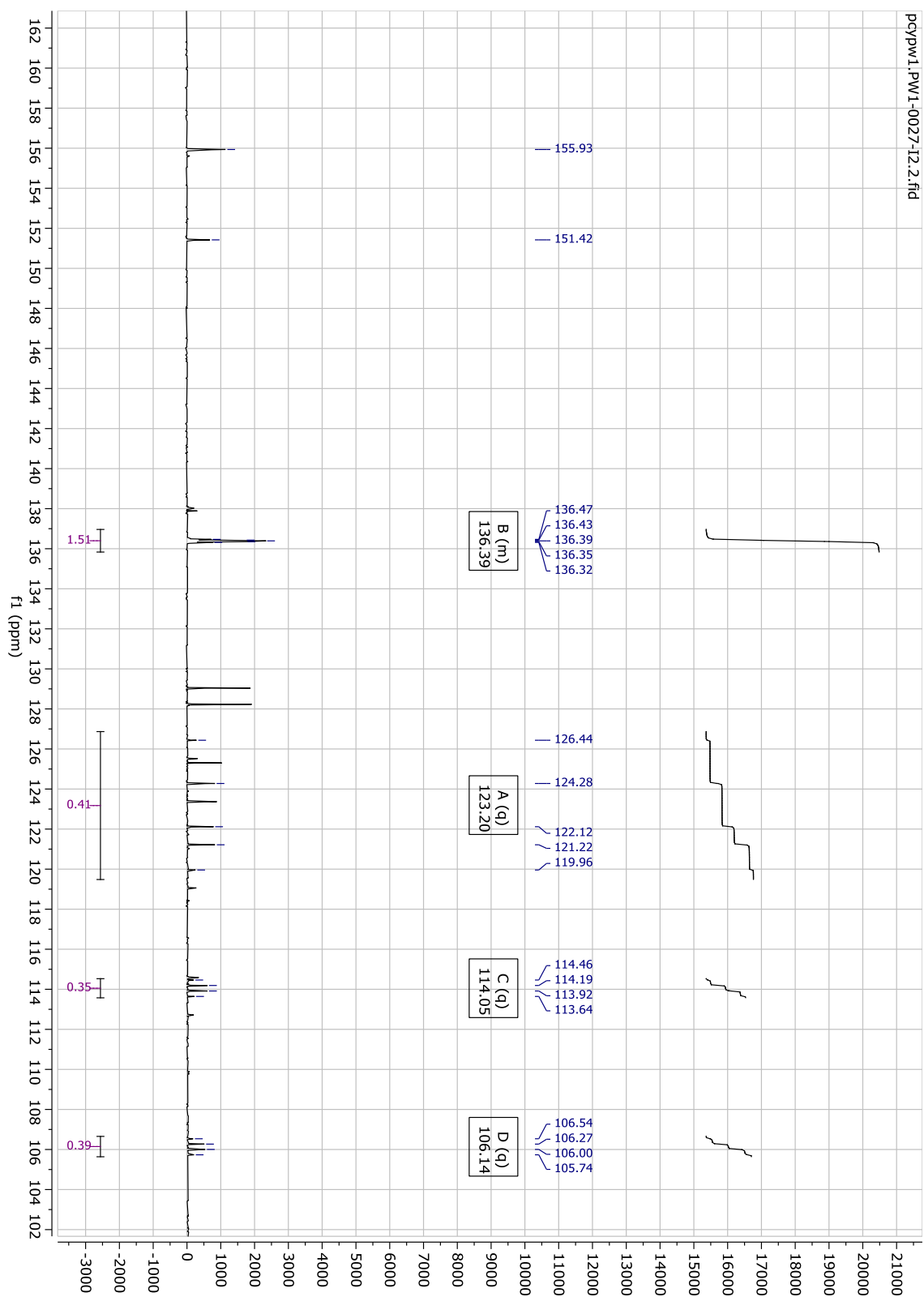


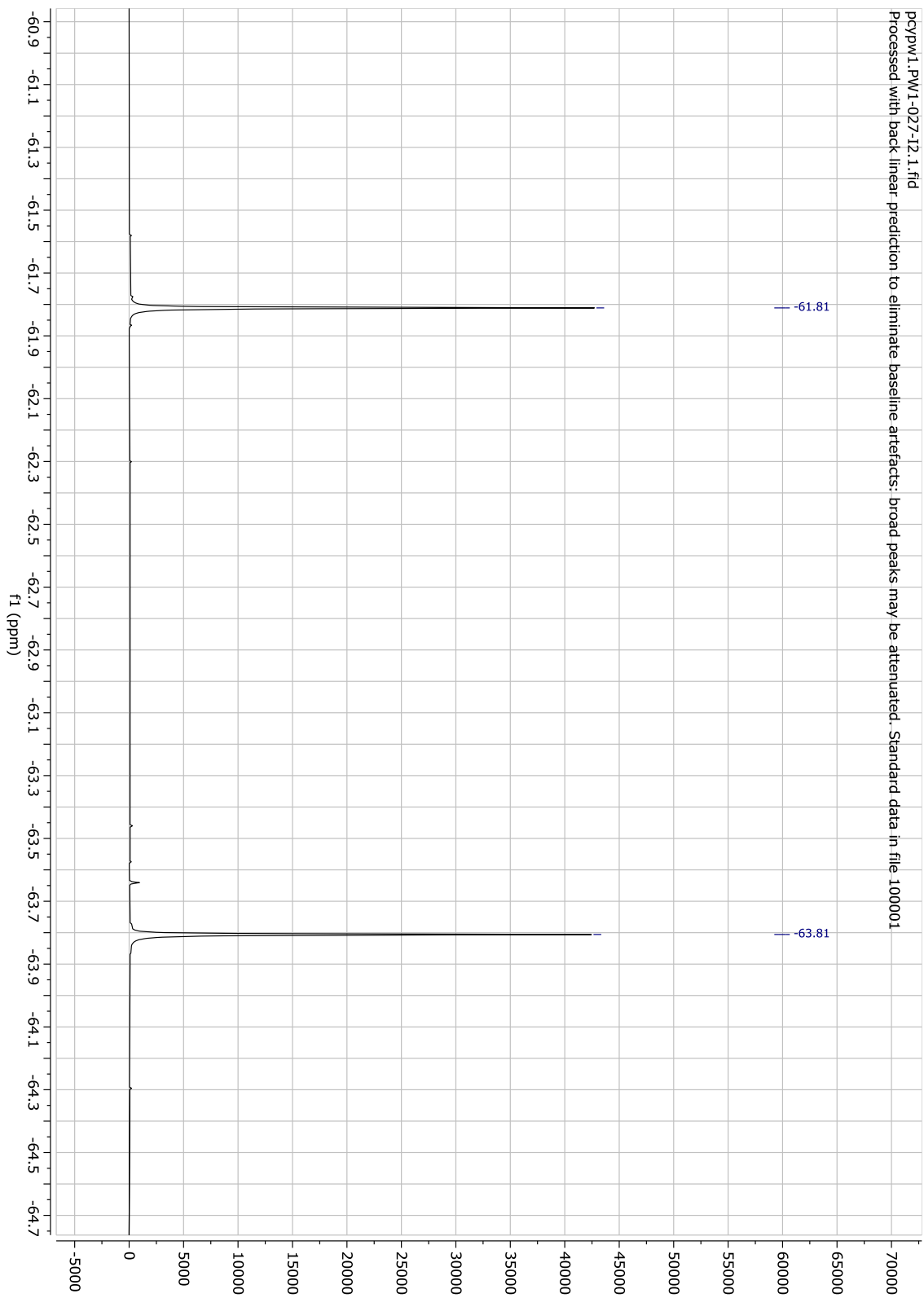
**87**

To a mixture of dichloromethane (35 mL) and water (15 mL) was added zinc trifluoromethanesulfinate (2.6 g, 7.8 mmol) and 6-chloropyridin-2-amine (0.5 g, 3.9 mmol). The mixture was cooled and *tert*-Butyl hydroperoxide (1.3 mL, 12 mmol) was added dropwise over the course of 5 minutes and the reaction mixture was heated at 50 °C for 48 hours. After this time, the reaction mixture was quenched with EDTA:sodium hydrogen carbonate (1:1 mixture of a 4 M aqueous solution and a saturated aqueous solution) (50 mL) and the organic layer dried over MgSO<sub>4</sub>, filtered and evaporated under reduced pressure. The residue was diluted with dichloromethane and adsorbed onto silica gel. Purification by silica gel chromatography, eluting with ethyl acetate, toluene and pentane (4.85:0.15:95), provided the *title compound* (26 mg, 2.5%) as a yellow solid: **mp.** 106-107 °C; **R<sub>f</sub>** 0.31 (10% EtOAc:90% Pentane); **FT-ATR**  $\nu_{\text{max}}$  3509, 3324, 3197, 1640, 1613, 1560, 1494, 1412, 1356, 1296, 1260, 1166, 1114, 1038, 965, 943, 778; **<sup>1</sup>H NMR** (500 MHz, CDCl<sub>3</sub>)  $\delta$  7.97 (s, 1H), 5.56 (s, 2H); **<sup>13</sup>C NMR** (126 MHz, CDCl<sub>3</sub>)  $\delta$  155.9, 151.4, 136.4 (hept,  $J$  = 4.7 Hz) (m), 123.2 (q,  $J$  = 271.6 Hz), 122.3 (q,  $J$  = 271.1 Hz), 114.1 (q,  $J$  = 34.4 Hz), 106.1 (q,  $J$  = 33.7 Hz); **<sup>19</sup>F NMR** (376 MHz, CDCl<sub>3</sub>)  $\delta$  -61.81, -63.81; **HRMS**  $m/z$  (ESI)<sup>-</sup> calcd for C<sub>7</sub>H<sub>2</sub>ClF<sub>6</sub>N<sub>2</sub> [M-H]<sup>-</sup> requires 262.9816, found 262.9823.

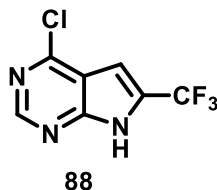




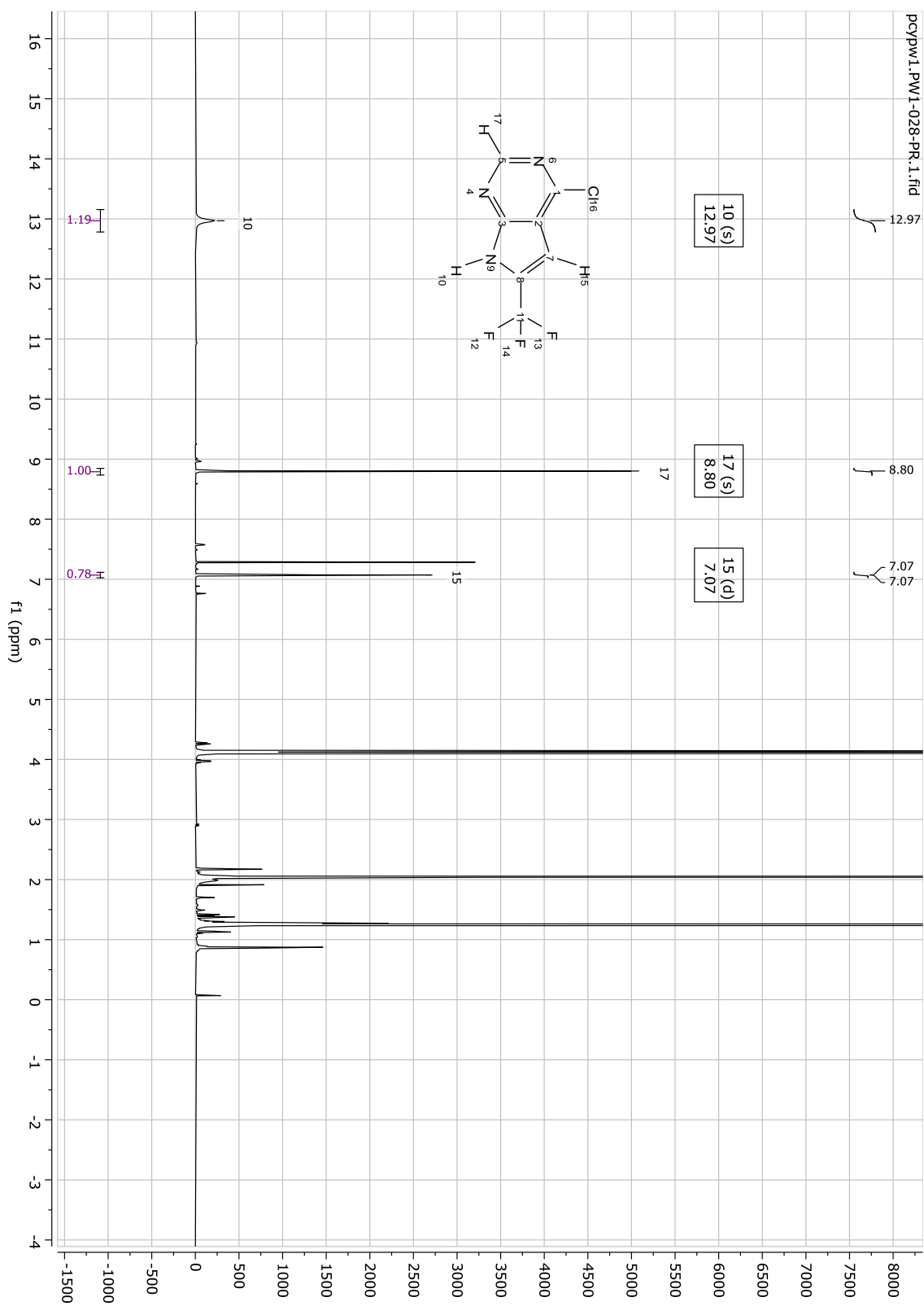


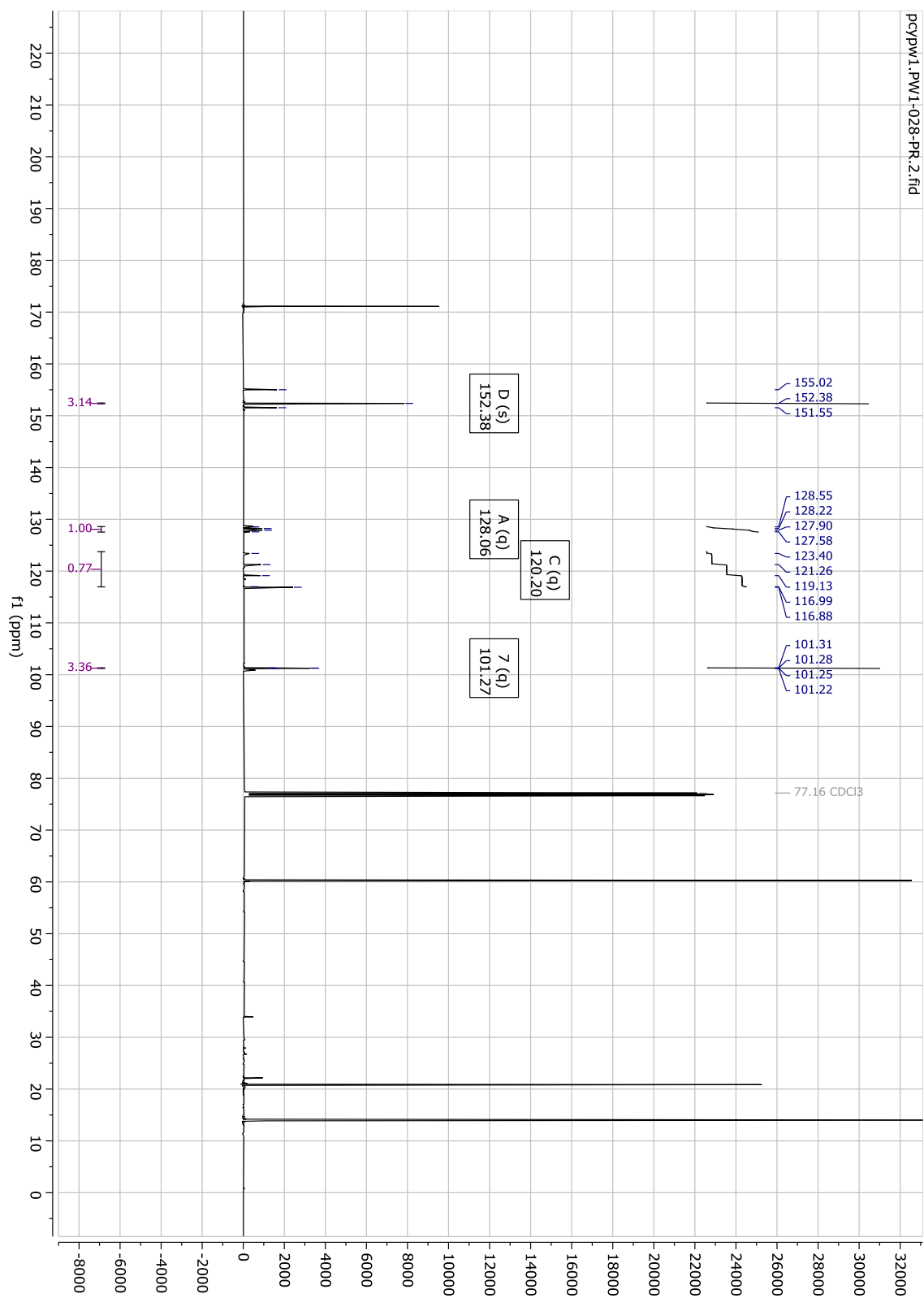


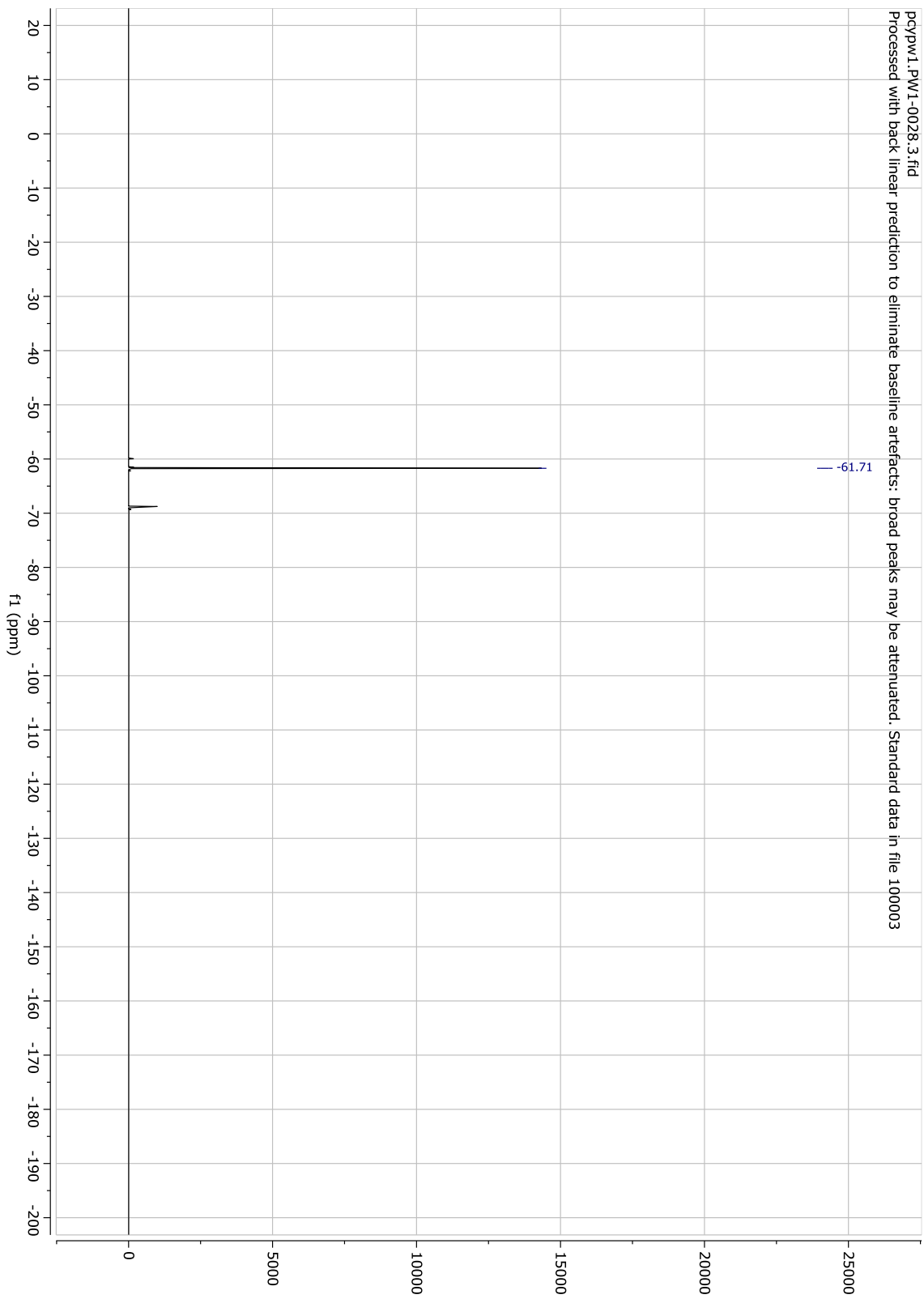
#### 4-chloro-6-(trifluoromethyl)-7H-pyrrolo[2,3-d]pyrimidine



To a solution of dichloromethane (14 mL) and water (6 mL) was added zinc trifluoromethanesulfinate (945.2 mg, 2.6 mmol) and 6-chloropyridin-2-amine (200.0 mg, 1.3 mmol). The mixture was cooled and *tert*-Butyl hydroperoxide (378  $\mu$ L, 1.2 mmol) was added dropwise before heating to 50 °C and left to stir for 48 h. Following this time, the reaction mixture was quenched with EDTA:sodium hydrogen carbonate (1:1 mixture of a 4 M aqueous solution and a saturated aqueous solution) (30 mL) and the organic layer dried over  $\text{MgSO}_4$ , filtered and evaporated under reduced pressure. The residue was diluted with dichloromethane and adsorbed onto silica gel. Purification by silica gel chromatography, eluting with ethyl acetate and pentane (5:95 to 10:90), provided the *title compound* (110 mg, 38%): **mp.** 187-189°C; **R<sub>f</sub>** 0.22 (10% EtOAc:90% Pentane); **FT-ATR**  $\nu_{\text{max}}$  3092, 2757, 1610, 1573, 1422, 1363, 1307, 1241, 1217, 1178, 1121, 985, 938, 830, 775, 751; **<sup>1</sup>H NMR** (500 MHz,  $\text{CDCl}_3$ )  $\delta$  12.97 (s, 1H), 8.80 (s, 1H), 7.07 (d,  $J$  = 1.3 Hz, 1H); **<sup>13</sup>C NMR** (126 MHz,  $\text{CDCl}_3$ )  $\delta$  155.0, 152.4, 151.6, 128.06 (q,  $J$  = 40.4 Hz), 120.2 (q,  $J$  = 268.8 Hz), 116.9, 101.3 (q,  $J$  = 3.7 Hz); **<sup>19</sup>F NMR** (376 MHz,  $\text{CDCl}_3$ )  $\delta$  -61.71; **HRMS**  $m/z$  (ESI)<sup>-</sup> calcd for  $\text{C}_7\text{H}_2\text{ClF}_3\text{N}_3$  [M-H]<sup>-</sup> requires 219.9895, found 219.9899.

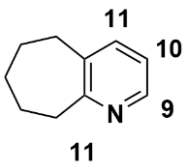
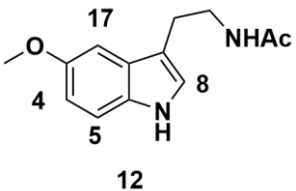
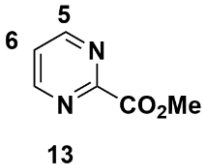
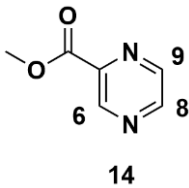
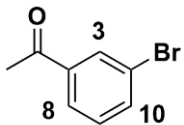




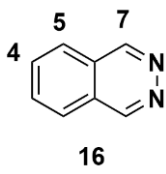
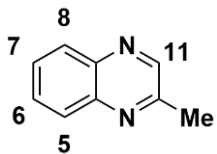
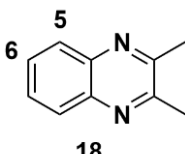
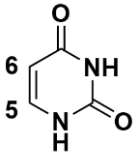
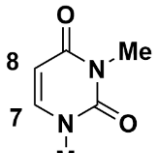


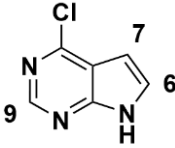
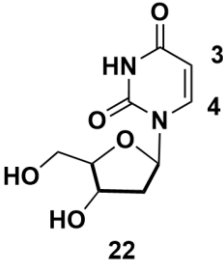
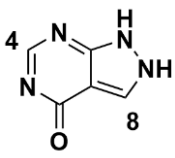
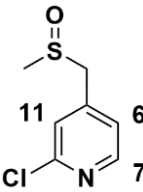
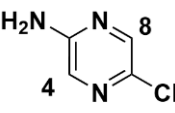
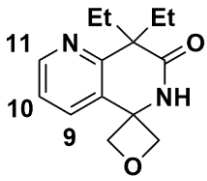
## 7.3 Activation Energy Calculation Data

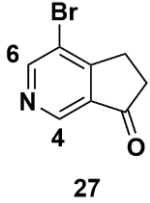
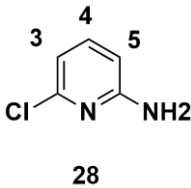
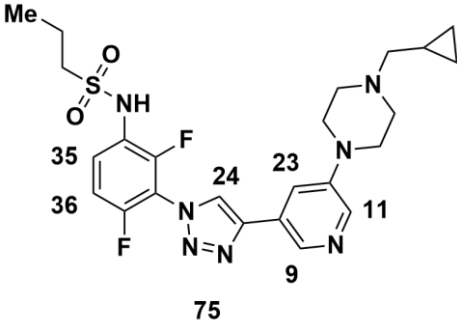
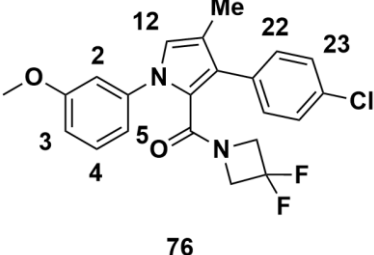
Calculated activation energies for experimentally observed compounds gathered from the literature.<sup>140</sup>

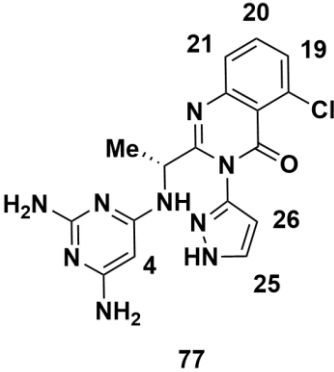
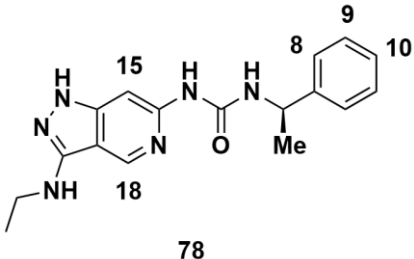
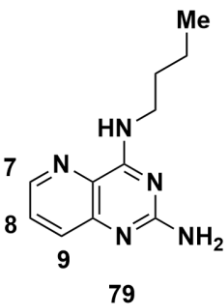
Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 11	9	Major	8.8
	10	Minor	9.1
	11		9.0
 12	4		7.1
	5		6.3
	8	Major	3.9
	17		5.4
 13	5	Major	12.0
	6	Major	11.4
 14	6		9.2
	8	Major	8.5
	9		9.4
 15	3		9.9
	8	Major	9.6
	10	Minor	9.8

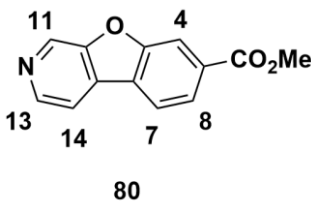
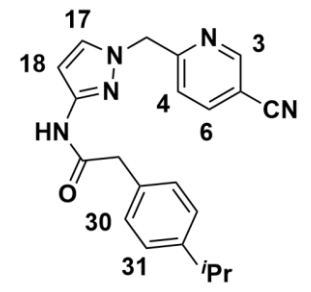
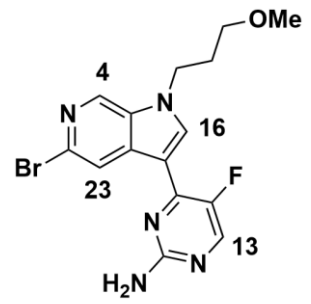


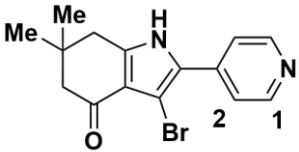
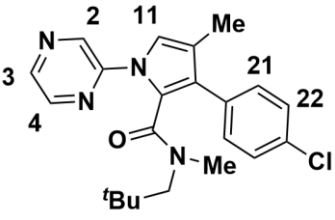
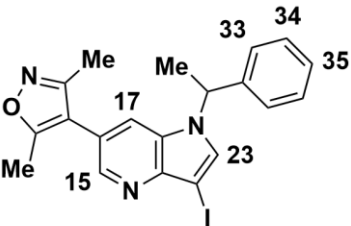
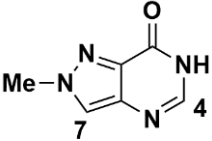
Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 16	4	Minor	7.7
	5	Major	6.9
	7		7.5
 17	5		7.3
	6		8.3
	7		8.1
	8		7.3
	11		8.1
 18	5	Major	6.3
	6	Minor	7.1
 19	5		10.9
	6	Major	10.9
 20	7		10.9
	8	Major	10.7

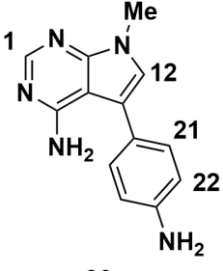
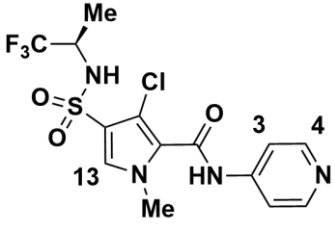
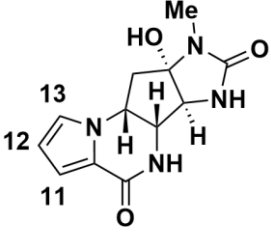
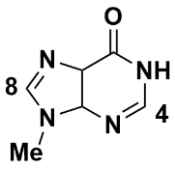
Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 <p>21</p>	6	Major	8.6
	7		12.4
	9		11.0
 <p>22</p>	3	Major	10.5
	4		12.1
 <p>23</p>	4		17.9
	8	Major	15.1
 <p>24</p>	6		10.6
	7	Major	9.9
	11		10.2
 <p>25</p>	4	Major	9.0
	8		10.0
 <p>26</p>	9		13.2
	10		10.7
	11	Major	10.4

Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 <p>27</p>	4	Major	8.4
	6		9.3
 <p>28</p>	3		10.0
	4		10.2
	5	Major	8.2
 <p>75</p>	9		N/A
	11	Major	N/A
	23		N/A
	24		N/A
	35		N/A
	36		N/A
 <p>76</p>	2		14.8
	3		9.6
	4		9.7
	5		10.1
	12	Major	10.0
	22		12.0
	23		9.0

Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 <p>77</p>	4	Major	20.2
	19		10.5
	20		10.8
	21		9.6
	25		15.0
	26		21.1
 <p>78</p>	8		10.0
	9		9.4
	10		9.5
	15	Major	10.9
	18		10.5
 <p>79</p>	7	Major	11.9
	8		13.0
	9		10.4

Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 <p>80</p>	4		9.6
	7		22.4
	8		10.9
	11	Major	11.6
	13		10.7
	14		9.4
 <p>81</p>	3	Major	10.4
	4		10.7
	6	Major	11.0
	17		13.1
	18		17.6
	30		9.6
	31		11.6
 <p>82</p>	4	Major	10.8
	13		13.7
	16		7.4
	23		10.3

Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 <p>86</p>	1	Major	9.8
	2		10.0
 <p>87</p>	2		11.2
	3		11.3
	4		10.0
	11	Major	10.0
	21		10.8
	22		11.1
 <p>88</p>	15		10.9
	17		10.3
	23	Major	9.1
	33		9.2
	34		9.9
	35		9.6
 <p>89</p>	4		12.5
	7	Major	10.7

Compound	Site	Experimentally Observed Products	HF Activation Energy (kcal mol <sup>-1</sup> )
 <p>90</p>	1		13.5
	12	Major	7.8
	21		10.6
	22		9.1
 <p>91</p>	3		10.3
	4	Major	10.6
	13		17.2
 <p>92</p>	11		14.3
	12		16.6
	13	Major	11.0
 <p>93</p>	4		13.1
	8	Major	12.7

## 7.4 Computational Workflow Details

### 7.4.1 Explanation on Z-matrices

Z matrices are a form of atom system representation whereby atom's positions are defined by their relationship to one other rather than points in Cartesian space. In this representation, the first atom is given position 0,0,0 and the second atom in the system has its position related to the first by bond length, bond angle and dihedral angle. For example, if the bond length between atom 1 and atom 2 is 1.8 Å the z-matrix for atom 2 is 1.8,0,0 1,0,0 where the first matrix gives the distances and angles and the second matrix is the atom number that the measurement is related to. Since there are not enough atoms in the above example to give the angle (matrix 1 position 2) and dihedral angle (matrix 1 position 3) these are left as 0 in both matrix elements 1 and 2.



## Chapter 8 - Bibliography

- (1) Kuntz, D.; Wilson, A. K. Machine Learning, Artificial Intelligence, and Chemistry: How Smart Algorithms Are Reshaping Simulation and the Laboratory. *Pure Appl. Chem.* **2022**, *94* (8), 1019–1054. <https://doi.org/10.1515/pac-2022-0202>.
- (2) *Global pharma companies' return on R&D investment increases after record low* | Deloitte UK. <https://www.deloitte.com/uk/en/about/press-room/global-pharma-companies-return-on-rd-investment-increases-after-record-low.html> (accessed 2024-10-07).
- (3) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic Synthesis Provides Opportunities to Transform Drug Discovery. *Nat Chem* **2018**, *10* (4), 383–394. <https://doi.org/10.1038/s41557-018-0021-z>.
- (4) Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Hou, T.; Song, M. Recent Advances in Deep Learning for Retrosynthesis. *WIREs Comput. Mol. Sci.* **2024**, *14* (1), e1694. <https://doi.org/10.1002/wcms.1694>.
- (5) Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture). *Angew. Chem., Int. Ed.* **1991**, *30* (5), 455–465. <https://doi.org/10.1002/anie.199104553>.
- (6) Corey, E. J. Retrosynthetic Thinking-Essentials and Examples. *Chem. Soc. Rev.* **1988**, *17*, 111–133.
- (7) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55* (20), 5904–5937. <https://doi.org/10.1002/anie.201506101>.
- (8) Feng, F.; Lai, L.; Pei, J. Computational Chemical Synthesis Analysis and Pathway Design. *Front Chem* **2018**, *6*.
- (9) Baskin, I. I.; Winkler, D.; Tetko, I. V. A Renaissance of Neural Networks in Drug Discovery. *Expert Opin Drug Discov* **2016**, *11* (8), 785–795. <https://doi.org/10.1080/17460441.2016.1201262>.

- (10) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov Today* **2018**, *23* (6), 1241–1250. <https://doi.org/10.1016/j.drudis.2018.01.039>.
- (11) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23* (25), 5966–5971. <https://doi.org/10.1002/chem.201605499>.
- (12) Heifets, A.; Jurisica, I. Construction of New Medicines via Game Proof Search. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *26* (1), 1564–1570. <https://doi.org/10.1609/aaai.v26i1.8331>.
- (13) Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. *ArXiv* **2020**, *abs/2001.01408*.
- (14) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. <https://doi.org/10.1038/nature25978>.
- (15) Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; Colton, S. A Survey of Monte Carlo Tree Search Methods. *IEEE Trans Comput Intell AI Games* **2012**, *4* (1), 1–43. <https://doi.org/10.1109/TCIAIG.2012.2186810>.
- (16) Zhang, Y.; He, X.; Gao, S.; Zhou, A.; Hao, H. Evolutionary Retrosynthetic Route Planning [Research Frontier]. *IEEE Comput Intell Mag* **2024**, *19* (3), 58–72. <https://doi.org/10.1109/MCI.2024.3401369>.
- (17) Akhmetshin, T.; Zankov, D.; Gantzer, P.; Babadeev, D.; Pinigina, A.; Madzhidov, T.; Varnek, A. SynPlanner: An End-to-End Tool for Synthesis Planning. *ChemRxiv*. August 6, 2024. <https://doi.org/10.26434/chemrxiv-2024-bzpdn-v2>.
- (18) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A Fast, Robust and Flexible Open-Source Software for Retrosynthetic Planning. *J Cheminform* **2020**, *12* (1), 70. <https://doi.org/10.1186/s13321-020-00472-1>.
- (19) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science (1979)* **2019**, *365* (6453), eaax1566. <https://doi.org/10.1126/science.aax1566>.

- (20) consortium, wwPDB. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res* **2019**, *47* (D1), D520–D528. <https://doi.org/10.1093/nar/gky949>.
- (21) Jaskolski, M.; Dauter, Z.; Wlodawer, A. A Brief History of Macromolecular Crystallography, Illustrated by a Family Tree and Its Nobel Fruits. *FEBS J* **2014**, *281* (18), 3985–4009. <https://doi.org/10.1111/febs.12796>.
- (22) Bai, X.; McMullan, G.; Scheres, S. H. W. How Cryo-EM Is Revolutionizing Structural Biology. *Trends Biochem Sci* **2015**, *40* (1), 49–57. <https://doi.org/10.1016/j.tibs.2014.10.005>.
- (23) Wüthrich, K. The Way to NMR Structures of Proteins. *Nat Struct Biol* **2001**, *8* (11), 923–925. <https://doi.org/10.1038/nsb1101-923>.
- (24) Thompson, M. C.; Yeates, T. O.; Rodriguez, J. A. Advances in Methods for Atomic Resolution Macromolecular Structure Determination. *F1000Res* **2020**, *9* (667). <https://doi.org/10.12688/f1000research.25097.1>.
- (25) Mitchell, A. L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M. R.; Kale, V.; Potter, S. C.; Richardson, L. J.; Sakharova, E.; Scheremetjew, M.; Korobeynikov, A.; Shlemov, A.; Kunyavskaya, O.; Lapidus, A.; Finn, R. D. MGnify: The Microbiome Analysis Resource in 2020. *Nucleic Acids Res* **2020**, *48* (D1), D570–D578. <https://doi.org/10.1093/nar/gkz1035>.
- (26) Steinegger, M.; Mirdita, M.; Söding, J. Protein-Level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold. *Nat Methods* **2019**, *16* (7), 603–606. <https://doi.org/10.1038/s41592-019-0437-4>.
- (27) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The Protein Folding Problem. *Annu Rev Biophys* **2008**, *37* (Volume 37, 2008), 289–316. <https://doi.org/10.1146/annurev.biophys.37.092707.153558>.
- (28) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* (1979) **1973**, *181* (4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>.
- (29) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Topf, M. Critical Assessment of Techniques for Protein Structure Prediction, Fifteenth Round. *CASP 15 Abstract Book* **2022**.
- (30) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Židek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.;

- Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, 577 (7792), 706–710. <https://doi.org/10.1038/s41586-019-1923-7>.
- (31) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **2017**, 13 (1), e1005324-.
  - (32) Zheng, W.; Li, Y.; Zhang, C.; Pearce, R.; Mortuza, S. M.; Zhang, Y. Deep-Learning Contact-Map Guided Protein Structure Prediction in CASP13. *Proteins: Struct., Funct., Bioinf.* **2019**, 87 (12), 1149–1164. <https://doi.org/10.1002/prot.25792>.
  - (33) Abriata, L. A.; Tamò, G. E.; Dal Peraro, M. A Further Leap of Improvement in Tertiary Structure Prediction in CASP13 Prompts New Routes for Future Assessments. *Proteins: Struct., Funct., Bioinf.* **2019**, 87 (12), 1100–1112. <https://doi.org/10.1002/prot.25787>.
  - (34) Pearce, R.; Zhang, Y. Deep Learning Techniques Have Significantly Impacted Protein Structure Prediction and Protein Design. *Curr Opin Struct Biol* **2021**, 68, 194–207. <https://doi.org/10.1016/j.sbi.2021.01.007>.
  - (35) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, 596 (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
  - (36) *Accelerating the race against antibiotic resistance - Google DeepMind.* [https://deepmind.google/discover/blog/accelerating-the-race-against-antibiotic-resistance/?\\_gl=1\\*17aqegh\\*\\_up\\*MQ..\\*\\_ga\\*MTQ1MDAxNzA3MC4xNzI4MjkwMzg0\\*\\_ga\\_LS8HVHCNQ0\\*MTcyODI5MDM4NC4xLjAuMTcyODI5MDQ5My4wLjAuMA..](https://deepmind.google/discover/blog/accelerating-the-race-against-antibiotic-resistance/?_gl=1*17aqegh*_up*MQ..*_ga*MTQ1MDAxNzA3MC4xNzI4MjkwMzg0*_ga_LS8HVHCNQ0*MTcyODI5MDM4NC4xLjAuMTcyODI5MDQ5My4wLjAuMA..) (accessed 2024-10-07).
  - (37) *Targeting early-onset Parkinson's with AI - Google DeepMind.* [https://deepmind.google/discover/blog/targeting-early-onset-parkinsons-with-ai/?\\_gl=1\\*5jyrul\\*\\_up\\*MQ..\\*\\_ga\\*MTQ1MDAxNzA3MC4xNzI4MjkwMzg0\\*\\_ga\\_LS8HVHCNQ0\\*MTcyODI5MDM4NC4xLjAuMTcyODI5MDQ5OS4wLjAuMA..](https://deepmind.google/discover/blog/targeting-early-onset-parkinsons-with-ai/?_gl=1*5jyrul*_up*MQ..*_ga*MTQ1MDAxNzA3MC4xNzI4MjkwMzg0*_ga_LS8HVHCNQ0*MTcyODI5MDM4NC4xLjAuMTcyODI5MDQ5OS4wLjAuMA..) (accessed 2024-10-07).

- (38) Valente, E. M.; Abou-Sleiman, P. M.; Caputo, V.; Muqit, M. M. K.; Harvey, K.; Gispert, S.; Ali, Z.; Del Turco, D.; Bentivoglio, A. R.; Healy, D. G.; Albanese, A.; Nussbaum, R.; González-Maldonado, R.; Deller, T.; Salvi, S.; Cortelli, P.; Gilks, W. P.; Latchman, D. S.; Harvey, R. J.; Dallapiccola, B.; Auburger, G.; Wood, N. W. Hereditary Early-Onset Parkinson's Disease Caused by Mutations in PINK1. *Science* (1979) **2004**, 304 (5674), 1158–1160. <https://doi.org/10.1126/SCIENCE.1096284>.
- (39) Ge, P.; Dawson, V. L.; Dawson, T. M. PINK1 and Parkin Mitochondrial Quality Control: A Source of Regional Vulnerability in Parkinson's Disease. *Mol Neurodegener* **2020**, 15 (1). <https://doi.org/10.1186/S13024-020-00367-7>.
- (40) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C. C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Židek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* 2024 630:8016 **2024**, 630 (8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>.
- (41) Barros, E. P.; Schiffer, J. M.; Vorobieva, A.; Dou, J.; Baker, D.; Amaro, R. E. Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics Simulations. *J Chem Theory Comput* **2019**, 15 (10), 5703–5715. <https://doi.org/10.1021/acs.jctc.9b00483>.
- (42) Bale, J. B.; Gonen, S.; Liu, Y.; Sheffler, W.; Ellis, D.; Thomas, C.; Cascio, D.; Yeates, T. O.; Gonen, T.; King, N. P.; Baker, D. Accurate Design of Megadalton-Scale Two-Component Icosahedral Protein Complexes. *Science* (1979) **2016**, 353 (6297), 389–394. <https://doi.org/10.1126/science.aaf8818>.
- (43) Langan, R. A.; Boyken, S. E.; Ng, A. H.; Samson, J. A.; Dods, G.; Westbrook, A. M.; Nguyen, T. H.; Lajoie, M. J.; Chen, Z.; Berger, S.; Mulligan, V. K.; Dueber, J. E.; Novak, W. R. P.; El-Samad, H.; Baker, D. De Novo Design of Bioactive Protein Switches. *Nature* **2019**, 572 (7768), 205–210. <https://doi.org/10.1038/s41586-019-1432-8>.
- (44) Bick, M. J.; Greisen, P. J.; Morey, K. J.; Antunes, M. S.; La, D.; Sankaran, B.; Reymond, L.; Johnsson, K.; Medford, J. I.; Baker, D. Computational Design of

Environmental Sensors for the Potent Opioid Fentanyl. *Elife* **2017**, *6*, e28909. <https://doi.org/10.7554/eLife.28909>.

- (45) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal* **2020**, *10* (3), 2260–2297. <https://doi.org/10.1021/ACSCATAL.9B04186>/ASSET/IMAGES/LARGE/CS9B04186\_0030.JPEG.
- (46) Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys Rev Lett* **2012**, *108* (5). <https://doi.org/10.1103/PHYSREVLETT.108.058301>.
- (47) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci Rep* **2013**, *3*. <https://doi.org/10.1038/SREP02810>.
- (48) Faber, F.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int J Quantum Chem* **2015**, *115* (16), 1094–1101. <https://doi.org/10.1002/QUA.24917>.
- (49) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys Rev B Condens Matter Mater Phys* **2014**, *89* (9). <https://doi.org/10.1103/PHYSREVB.89.094104>.
- (50) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-Aided Bandgap Engineering for Solar Materials. *Comput Mater Sci* **2014**, *83*, 185–195. <https://doi.org/10.1016/j.commatsci.2013.10.016>.
- (51) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE Journal* **2018**, *64* (7), 2311–2323. <https://doi.org/10.1002/aic.16198>.
- (52) Hattori, T.; Kito, S. Artificial Intelligence Approach to Catalyst Design. *Catal Today* **1991**, *10* (2), 213–222. [https://doi.org/10.1016/0920-5861\(91\)80066-I](https://doi.org/10.1016/0920-5861(91)80066-I).
- (53) Hattori, T.; Kito, S.; Murakami, Y. Integration of Catalyst Activity Pattern (INCAP) Artificial Intelligence Approach in Catalyst Design. *Chem Lett* **1988**, *17* (8), 1269–1272. <https://doi.org/10.1246/cl.1988.1269>.

- (54) Hattori, T.; Kito, S. Neural Network as a Tool for Catalyst Development. *Catal Today* **1995**, *23* (4), 347–355. [https://doi.org/10.1016/0920-5861\(94\)00148-U](https://doi.org/10.1016/0920-5861(94)00148-U).
- (55) Suzuki, K.; Toyao, T.; Maeno, Z.; Takakusagi, S.; Shimizu, K.; Takigawa, I. Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data. *ChemCatChem* **2019**, *11* (18), 4537–4547. <https://doi.org/10.1002/cctc.201900971>.
- (56) Omata, K. Screening of New Additives of Active-Carbon-Supported Heteropoly Acid Catalyst for Friedel–Crafts Reaction by Gaussian Process Regression. *Ind Eng Chem Res* **2011**, *50* (19), 10948–10954. <https://doi.org/10.1021/ie102477y>.
- (57) Can, E.; Yildirim, R. Data Mining in Photocatalytic Water Splitting over Perovskites Literature for Higher Hydrogen Production. *Appl Catal B* **2019**, *242*, 267–283. <https://doi.org/10.1016/j.apcatb.2018.09.104>.
- (58) Rydberg, P.; Rostkowski, M.; Gloriam, D. E.; Olsen, L. The Contribution of Atom Accessibility to Site of Metabolism Models for Cytochromes P450. *Mol Pharm* **2013**, *10* (4), 1216–1223. <https://doi.org/10.1021/mp3005116>.
- (59) Rydberg, P.; Gloriam, D. E.; Zaretski, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med Chem Lett* **2010**, *1* (3), 96–100. <https://doi.org/10.1021/ml100016x>.
- (60) Olsen, L.; Montefiori, M.; Tran, K. P.; Jørgensen, F. S. SMARTCyp 3.0: Enhanced Cytochrome P450 Site-of-Metabolism Prediction Server. *Bioinformatics* **2019**, *35* (17), 3174–3175. <https://doi.org/10.1093/bioinformatics/btz037>.
- (61) Rydberg, P.; Rostkowski, M.; Gloriam, D. E.; Olsen, L. The Contribution of Atom Accessibility to Site of Metabolism Models for Cytochromes P450. *Mol. Pharm.* **2013**, *10* (4), 1216–1223. <https://doi.org/10.1021/mp3005116>.
- (62) Öeren, M.; Walton, P. J.; Hunt, P. A.; Ponting, D. J.; Segall, M. D. Predicting Reactivity to Drug Metabolism: Beyond P450s—Modelling FMOs and UGTs. *J Comput Aided Mol Des* **2021**, *35* (4), 541–555. <https://doi.org/10.1007/S10822-020-00321-1/FIGURES/17>.
- (63) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11* (6), 2336–2347. [https://doi.org/10.1021/ACS.JPCLETT.9B03664/SUPPL\\_FILE/JZ9B03664\\_LIVE\\_SLIDES.MP4](https://doi.org/10.1021/ACS.JPCLETT.9B03664/SUPPL_FILE/JZ9B03664_LIVE_SLIDES.MP4).

- (64) Krems, R. V. Bayesian Machine Learning for Quantum Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2019**, *21* (25), 13392–13410. <https://doi.org/10.1039/C9CP01883B>.
- (65) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural Network Models of Potential Energy Surfaces. *J. Chem. Phys.* **1995**, *103* (10), 4129–4137. <https://doi.org/10.1063/1.469597>.
- (66) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-Kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Comput Phys Commun* **2018**, *228*, 178–184. <https://doi.org/10.1016/J.CPC.2018.03.016>.
- (67) Jinnouchi, R.; Karsai, F.; Kresse, G. On-the-Fly Machine Learning Force Field Generation: Application to Melting Points. *Phys Rev B* **2019**, *100* (1). <https://doi.org/10.1103/PHYSREVB.100.014105>.
- (68) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K. R. SchNetPack: A Deep Learning Toolbox for Atomistic Systems. *J Chem Theory Comput* **2019**, *15* (1), 448–455. <https://doi.org/10.1021/ACS.JCTC.8B00908>.
- (69) Westermayr, J.; Gastegger, M.; Menger, M. F. S. J.; Mai, S.; González, L.; Marquetand, P. Machine Learning Enables Long Time Scale Molecular Photodynamics Simulations. *Chem Sci* **2019**, *10* (35), 8100–8107. <https://doi.org/10.1039/C9SC01742A>.
- (70) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148* (24). <https://doi.org/10.1063/1.5019779>.
- (71) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *J. Phys. Chem. Lett.* **2018**, *9* (11), 2725–2732. <https://doi.org/10.1021/ACS.JPCLETT.8B00684>.
- (72) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys Rev Lett* **2018**, *120* (14). <https://doi.org/10.1103/PHYSREVLETT.120.143001>.
- (73) Chmiela, S.; Sauceda, H. E.; Müller, K. R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *Nat. Commun.* **2018**, *9* (1). <https://doi.org/10.1038/S41467-018-06169-2>.



- (74) Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. *Int J Quantum Chem* **2015**, *115* (16), 1074–1083. <https://doi.org/10.1002/QUA.24836>.
- (75) Handley, C. M.; Popelier, P. L. A. Potential Energy Surfaces Fitted by Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114* (10), 3371–3383. <https://doi.org/10.1021/JP9105585>.
- (76) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121* (1), 511–522. <https://doi.org/10.1021/ACS.JPCC.6B10908>.
- (77) Li, P.; Jia, X.; Pan, X.; Shao, Y.; Mei, Y. Accelerated Computation of Free Energy Profile at Ab Initio Quantum Mechanical/Molecular Mechanics Accuracy via a Semi-Empirical Reference Potential. I. Weighted Thermodynamics Perturbation. *J Chem Theory Comput* **2018**, *14* (11), 5583–5596. <https://doi.org/10.1021/ACS.JCTC.8B00571>.
- (78) Yao, K.; Herr, J. E.; Toth, D. W.; McKintyre, R.; Parkhill, J. The TensorMol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics. *Chem Sci* **2018**, *9* (8), 2261–2269. <https://doi.org/10.1039/C7SC04934J>.
- (79) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145* (17). <https://doi.org/10.1063/1.4966192>.
- (80) Shen, L.; Yang, W. Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. *J Chem Theory Comput* **2018**, *14* (3), 1442–1455. <https://doi.org/10.1021/ACS.JCTC.7B01195>.
- (81) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K. R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci Adv* **2017**, *3* (5). <https://doi.org/10.1126/SCIADV.1603015>.
- (82) Behler, J. Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations. *Phys. Chem. Chem. Phys.* **2011**, *13* (40), 17930–17955. <https://doi.org/10.1039/C1CP21668F>.
- (83) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys Rev Lett* **2015**, *114* (9). <https://doi.org/10.1103/PHYSREVLETT.114.096405>.

- (84) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys Rev Lett* **2007**, *98* (14). <https://doi.org/10.1103/PHYSREVLETT.98.146401>.
- (85) Ludwig, J.; Vlachos, D. G. Ab Initio Molecular Dynamics of Hydrogen Dissociation on Metal Surfaces Using Neural Networks and Novelty Sampling. *J. Chem. Phys.* **2007**, *127* (15). <https://doi.org/10.1063/1.2794338>.
- (86) Zheng, X.; Hu, L.; Wang, X.; Chen, G. A Generalized Exchange-Correlation Functional: The Neural-Networks Approach. *Chem Phys Lett* **2004**, *390* (1), 186–192. <https://doi.org/10.1016/j.cplett.2004.04.020>.
- (87) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the Electronic Structure Problem with Machine Learning. *npj Comput Mater* **2019**, *5* (1), 1–7. <https://doi.org/10.1038/s41524-019-0162-7>.
- (88) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K. R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8* (1), 1–10. <https://doi.org/10.1038/s41467-017-00839-3>.
- (89) Mueller, T.; Hernandez, A.; Wang, C. Machine Learning for Interatomic Potential Models. *J. Chem. Phys.* **2020**, *152* (5). <https://doi.org/10.1063/1.5126336>.
- (90) Eshet, H.; Khaliullin, R. Z.; Kühne, T. D.; Behler, J.; Parrinello, M. Ab Initio Quality Neural-Network Potential for Sodium. *Phys Rev B Condens Matter Mater Phys* **2010**, *81* (18). <https://doi.org/10.1103/PHYSREVB.81.184107>.
- (91) Le, H. M.; Huynh, S.; Raff, L. M. Molecular Dissociation of Hydrogen Peroxide (HOOH) on a Neural Network Ab Initio Potential Surface with a New Configuration Sampling Method Involving Gradient Fitting. *J. Chem. Phys.* **2009**, *131* (1). <https://doi.org/10.1063/1.3159748>.
- (92) Chen, W. K.; Liu, X. Y.; Fang, W. H.; Dral, P. O.; Cui, G. Deep Learning for Nonadiabatic Excited-State Dynamics. *J. Phys. Chem. Lett.* **2018**, *9* (23), 6702–6708. <https://doi.org/10.1021/ACS.JPCLETT.8B03026>.
- (93) Le, H. M.; Raff, L. M. Molecular Dynamics Investigation of the Bimolecular Reaction  $\text{BeH} + \text{H}_2 \rightarrow \text{BeH}_2 + \text{H}$  on an Ab Initio Potential-Energy Surface Obtained Using Neural Network Methods with Both Potential and Gradient Accuracy Determination. *J. Phys. Chem. A* **2010**, *114* (1), 45–53. <https://doi.org/10.1021/JP907507Z>.

- (94) Jindal, S.; Bulusu, S. S. A Transferable Artificial Neural Network Model for Atomic Forces in Nanoparticles. *J. Chem. Phys.* **2018**, *149* (19). <https://doi.org/10.1063/1.5043247>.
- (95) Agrawal, P. M.; Raff, L. M.; Hagan, M. T.; Komanduri, R. Molecular Dynamics Investigations of the Dissociation of Si O<sub>2</sub> on an Ab Initio Potential Energy Surface Obtained Using Neural Network Methods. *J. Chem. Phys.* **2006**, *124* (13). <https://doi.org/10.1063/1.2185638>.
- (96) Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys Rev Lett* **2008**, *100* (18). <https://doi.org/10.1103/PHYSREVLETT.100.185501>.
- (97) Brown, D. F. R.; Gibbs, M. N.; Clary, D. C. Combining Ab Initio Computations, Neural Networks, and Diffusion Monte Carlo: An Efficient Method to Treat Weakly Bound Molecules. *J. Chem. Phys.* **1996**, *105* (17), 7597–7604. <https://doi.org/10.1063/1.472596>.
- (98) Suzuki, T.; Tamura, R.; Miyazaki, T. Machine Learning for Atomic Forces in a Crystalline Solid: Transferability to Various Temperatures. *Int J Quantum Chem* **2017**, *117* (1), 33–39. <https://doi.org/10.1002/QUA.25307>.
- (99) Botu, V.; Ramprasad, R. Learning Scheme to Predict Atomic Forces and Accelerate Materials Simulations. *Phys Rev B Condens Matter Mater Phys* **2015**, *92* (9). <https://doi.org/10.1103/PHYSREVB.92.094306>.
- (100) Shilpa, S.; Kashyap, G.; Sunoj, R. B. Recent Applications of Machine Learning in Molecular Property and Chemical Reaction Outcome Predictions. *J. Phys. Chem. A* **2023**, *127* (40), 8253–8271. [https://doi.org/10.1021/ACS.JPCA.3C04779/ASSET/IMAGES/LARGE/JP3C04779\\_0003.JPEG](https://doi.org/10.1021/ACS.JPCA.3C04779/ASSET/IMAGES/LARGE/JP3C04779_0003.JPEG).
- (101) Madan, A. K.; Dureja, H. Prediction of Pharmacokinetic Parameters. **2012**, 337–357. [https://doi.org/10.1007/978-1-62703-050-2\\_14](https://doi.org/10.1007/978-1-62703-050-2_14).
- (102) Dureja, H.; Gupta, S.; Madan, A. K. Topological Models for Prediction of Physico-Chemical, Pharmacokinetic and Toxicological Properties of Antihistaminic Drugs Using Decision Tree and Moving Average Analysis. *Int J Comput Biol Drug Des* **2009**, *2* (4), 353–370. <https://doi.org/10.1504/IJCBDD.2009.030766>.
- (103) Dureja, H.; Gupta, S.; Madan, A. K. Topological Models for Prediction of Pharmacokinetic Parameters of Cephalosporins Using Random Forest,

Decision Tree and Moving Average Analysis. *Sci Pharm* **2008**, 76 (3), 377–394. <https://doi.org/10.3797/SCIPHARM.0803-30>.

- (104) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* (1979) **2018**, 360 (6385), 186–190. <https://doi.org/10.1126/SCIENCE.AAR5169>.
- (105) Sato, A.; Miyao, T.; Funatsu, K. Prediction of Reaction Yield for Buchwald–Hartwig Cross-Coupling Reactions Using Deep Learning. *Mol Inform* **2022**, 41 (2), 2100156. <https://doi.org/10.1002/minf.202100156>.
- (106) Zhao, W.; Li, Y. Predicting the Yield of Pd-Catalyzed Buchwald–Hartwig Amination Using Machine Learning with Extended Molecular Fingerprints and Selected Physical Parameters. *ChemistrySelect* **2024**, 9 (33), e202402529. <https://doi.org/10.1002/slct.202402529>.
- (107) Haywood, A. L.; Redshaw, J.; Hanson-Heine, M. W. D.; Taylor, A.; Brown, A.; Mason, A. M.; Gärtner, T.; Hirst, J. D. Kernel Methods for Predicting Yields of Chemical Reactions. *J Chem Inf Model* **2022**, 62 (9), 2077–2092. <https://doi.org/10.1021/acs.jcim.1c00699>.
- (108) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P. O. Organic Reactivity from Mechanism to Machine Learning. *Nat. Rev. Chem.* **2021**, 5 (4), 240–255. <https://doi.org/10.1038/s41570-021-00260-x>.
- (109) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuc, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, 7 (1). <https://doi.org/10.1038/s41598-017-02303-0>.
- (110) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, 6 (6), 1379–1390. <https://doi.org/10.1016/j.chempr.2020.02.017>.
- (111) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, 3 (5), 434–443. <https://doi.org/10.1021/acscentsci.7b00064>.
- (112) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; Desjarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial

Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63* (16), 8667–8682. <https://doi.org/10.1021/acs.jmedchem.9b02120>.

- (113) Engkvist, O.; Norrby, P. O.; Selmi, N.; Lam, Y. hong; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discov. Today* **2018**, *23* (6), 1203–1218. <https://doi.org/10.1016/j.drudis.2018.02.014>.
- (114) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725–732. <https://doi.org/10.1021/acscentsci.6b00219>.
- (115) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. <https://doi.org/10.1038/s41586-019-1384-z>.
- (116) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58* (14), 4515–4519. <https://doi.org/10.1002/anie.201806920>.
- (117) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. Fast and Accurate Prediction of the Regioselectivity of Electrophilic Aromatic Substitution Reactions. *Chem. Sci.* **2018**, *9* (3), 660–665. <https://doi.org/10.1039/c7sc04156j>.
- (118) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and on-the-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12* (6), 2198–2208. <https://doi.org/10.1039/d0sc04823b>.
- (119) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C–H Functionalization of Heterocycles through Machine Learning. *Angew. Chem., Int. Ed.* **2020**, *59* (32), 13253–13259. <https://doi.org/10.1002/anie.202000959>.
- (120) Yamaguchi, J.; Yamaguchi, A. D.; Itami, K. C–H Bond Functionalization: Emerging Synthetic Tools for Natural Products and Pharmaceuticals. *Angew. Chem., Int. Ed.* **2012**, *51* (36), 8960–9009. <https://doi.org/10.1002/anie.201201666>.

- (121) McMurray, L.; O'Hara, F.; Gaunt, M. J. Recent Developments in Natural Product Synthesis Using Metal-Catalysed C–H Bond Functionalisation. *Chem. Soc. Rev.* **2011**, *40* (4), 1885–1898. <https://doi.org/10.1039/C1CS15013H>.
- (122) Newhouse, T.; Baran, P. S. If C-H Bonds Could Talk: Selective C-H Bond Oxidation. *Angew. Chem., Int. Ed.* **2011**, *50* (15), 3362–3374. <https://doi.org/10.1002/anie.201006368>.
- (123) Gutekunst, W. R.; Baran, P. S. C–H Functionalization Logic in Total Synthesis. *Chem. Soc. Rev.* **2011**, *40* (4), 1976–1991. <https://doi.org/10.1039/C0CS00182A>.
- (124) Ji, Y.; Brueckl, T.; Baxter, R. D.; Fujiwara, Y.; Seiple, I. B.; Su, S.; Blackmond, D. G.; Baran, P. S. Innate C-H Trifluoromethylation of Heterocycles. *Proc Natl Acad Sci U S A* **2011**, *108* (35), 14411–14415. <https://doi.org/10.1073/pnas.1109059108>.
- (125) Arndtsen, B. A.; Bergman, R. G.; Mobley, T. A.; Peterson, T. H. *Selective Intermolecular Carbon-Hydrogen Bond Activation by Synthetic Metal Complexes in Homogeneous Solution*; 1995; Vol. 28. <https://pubs.acs.org/sharingguidelines>.
- (126) Guillemard, L.; Kaplaneris, N.; Ackermann, L.; Johansson, M. J. Late-Stage C–H Functionalization Offers New Opportunities in Drug Discovery. *Nat. Rev. Chem.* Nature Research August 1, 2021, pp 522–545. <https://doi.org/10.1038/s41570-021-00300-6>.
- (127) Friis, S. D.; Johansson, M. J.; Ackermann, L. Cobalt-Catalysed C–H Methylation for Late-Stage Drug Diversification. *Nat Chem* **2020**, *12* (6), 511–519. <https://doi.org/10.1038/s41557-020-0475-7>.
- (128) Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W. The Medicinal Chemist's Toolbox for Late Stage Functionalization of Drug-like Molecules. *Chem. Soc. Rev.* **2016**, *45* (3), 546–576. <https://doi.org/10.1039/C5CS00628G>.
- (129) Taylor, A. P.; Robinson, R. P.; Fobian, Y. M.; Blakemore, D. C.; Jones, L. H.; Fadeyi, O. Modern Advances in Heterocyclic Chemistry in Drug Discovery. *Org Biomol Chem* **2016**, *14* (28), 6611–6637. <https://doi.org/10.1039/C6OB00936K>.
- (130) Sambiagio, C.; Schönbauer, D.; Blicke, R.; Dao-Huy, T.; Pototschnig, G.; Schaaf, P.; Wiesinger, T.; Zia, M. F.; Wencel-Delord, J.; Besset, T.; Maes, B. U. W.; Schnürch, M. A Comprehensive Overview of Directing Groups Applied in Metal-Catalysed C-H Functionalisation Chemistry. *Chem. Soc. Rev.* Royal

Society of Chemistry September 7, 2018, pp 6603–6743.  
<https://doi.org/10.1039/c8cs00201k>.

- (131) Dai, H.-X.; Stepan, A. F.; Plummer, M. S.; Zhang, Y.-H.; Yu, J.-Q. Divergent C–H Functionalizations Directed by Sulfonamide Pharmacophores: Late-Stage Diversification as a Tool for Drug Discovery. *J Am Chem Soc* **2011**, *133* (18), 7222–7228. <https://doi.org/10.1021/ja201708f>.
- (132) Friis, S. D.; Johansson, M. J.; Ackermann, L. Cobalt-Catalysed C–H Methylation for Late-Stage Drug Diversification. *Nat Chem* **2020**, *12* (6), 511–519. <https://doi.org/10.1038/s41557-020-0475-7>.
- (133) Hilton, M. C.; Dolewski, R. D.; McNally, A. Selective Functionalization of Pyridines via Heterocyclic Phosphonium Salts. *J Am Chem Soc* **2016**, *138* (42), 13806–13809. <https://doi.org/10.1021/jacs.6b08662>.
- (134) Rodriguez, R. A.; Pan, C.-M.; Yabe, Y.; Kawamata, Y.; Eastgate, M. D.; Baran, P. S. Palau'chlor: A Practical and Reactive Chlorinating Reagent. *J Am Chem Soc* **2014**, *136* (19), 6908–6911. <https://doi.org/10.1021/ja5031744>.
- (135) He, Z.-T.; Li, H.; Haydl, A. M.; Whiteker, G. T.; Hartwig, J. F. Trimethylphosphate as a Methylating Agent for Cross Coupling: A Slow-Release Mechanism for the Methylation of Arylboronic Esters. *J Am Chem Soc* **2018**, *140* (49), 17197–17202. <https://doi.org/10.1021/jacs.8b10076>.
- (136) Shi, H.; Lu, Y.; Weng, J.; Bay, K. L.; Chen, X.; Tanaka, K.; Verma, P.; Houk, K. N.; Yu, J.-Q. Differentiation and Functionalization of Remote C–H Bonds in Adjacent Positions. *Nat Chem* **2020**, *12* (4), 399–404. <https://doi.org/10.1038/s41557-020-0424-5>.
- (137) Foo, K.; Sella, E.; Thomé, I.; Eastgate, M. D.; Baran, P. S. A Mild, Ferrocene-Catalyzed C–H Imidation of (Hetero)Arenes. *J Am Chem Soc* **2014**, *136* (14), 5279–5282. <https://doi.org/10.1021/ja501879c>.
- (138) Langlois, B. R.; Laurent, E.; Roidot, N. Trifluoromethylation of Aromatic Compounds with Sodium Trifluoromethanesulfinate under Oxidative Conditions. *Tetrahedron Lett* **1991**, *32* (51), 7525–7528. [https://doi.org/10.1016/0040-4039\(91\)80524-A](https://doi.org/10.1016/0040-4039(91)80524-A).
- (139) Fujiwara, Y.; Dixon, J. A.; Rodriguez, R. A.; Baxter, R. D.; Dixon, D. D.; Collins, M. R.; Blackmond, D. G.; Baran, P. S. A New Reagent for Direct Difluoromethylation. *J Am Chem Soc* **2012**, *134* (3), 1494–1497. <https://doi.org/10.1021/ja211422g>.

- (140) Smith, J. M.; Dixon, J. A.; DeGruyter, J. N.; Baran, P. S. Alkyl Sulfinates: Radical Precursors Enabling Drug Discovery. *J Med Chem* **2019**, *62* (5), 2256–2264. <https://doi.org/10.1021/acs.jmedchem.8b01303>.
- (141) Yan, M.; Lo, J. C.; Edwards, J. T.; Baran, P. S. Radicals: Reactive Intermediates with Translational Potential. *J Am Chem Soc* **2016**, *138* (39), 12692–12714. <https://doi.org/10.1021/jacs.6b08856>.
- (142) Fujiwara, Y.; Dixon, J. A.; O'Hara, F.; Funder, E. D.; Dixon, D. D.; Rodriguez, R. A.; Baxter, R. D.; Herlé, B.; Sach, N.; Collins, M. R.; Ishihara, Y.; Baran, P. S. Practical and Innate Carbon–Hydrogen Functionalization of Heterocycles. *Nature* **2012**, *492* (7427), 95–99. <https://doi.org/10.1038/nature11680>.
- (143) Zhou, Q.; Ruffoni, A.; Gianatassio, R.; Fujiwara, Y.; Sella, E.; Shabat, D.; Baran, P. S. Direct Synthesis of Fluorinated Heteroarylether Bioisosteres. *Angew. Chem., Int. Ed.* **2013**, *52* (14), 3949–3952. <https://doi.org/10.1002/anie.201300763>.
- (144) Gui, J.; Zhou, Q.; Pan, C.-M.; Yabe, Y.; Burns, A. C.; Collins, M. R.; Ornelas, M. A.; Ishihara, Y.; Baran, P. S. C–H Methylation of Heteroarenes Inspired by Radical SAM Methyl Transferase. *J Am Chem Soc* **2014**, *136* (13), 4853–4856. <https://doi.org/10.1021/ja5007838>.
- (145) O'Hara, F.; Baxter, R. D.; O'Brien, A. G.; Collins, M. R.; Dixon, J. A.; Fujiwara, Y.; Ishihara, Y.; Baran, P. S. Preparation and Purification of Zinc Sulfinates Reagents for Drug Discovery. *Nat Protoc* **2013**, *8* (6), 1042–1047. <https://doi.org/10.1038/nprot.2013.059>.
- (146) Gianatassio, R.; Kawamura, S.; Eprile, C. L.; Foo, K.; Ge, J.; Burns, A. C.; Collins, M. R.; Baran, P. S. Simple Sulfinates Synthesis Enables C–H Trifluoromethylcyclopropanation. *Angew. Chem., Int. Ed.* **2014**, *53* (37), 9851–9855. <https://doi.org/10.1002/anie.201406622>.
- (147) Minisci, F.; Vismara, E.; Fontana, F. Recent Developments of Free-Radical Substitutions of Heteroaromatic Bases. *Heterocycles* **1989**, *28* (1), 489. <https://doi.org/10.3987/REV-88-SR1>.
- (148) Duncanson, M. A. J. Minisci Reactions: Versatile CH-Functionalizations for Medicinal Chemists. *Medchemcomm* **2011**, *2* (12), 1135–1161. <https://doi.org/10.1039/C1MD00134E>.
- (149) Seiple, I. B.; Su, S.; Rodriguez, R. A.; Gianatassio, R.; Fujiwara, Y.; Sobel, A. L.; Baran, P. S. Direct C–H Arylation of Electron-Deficient Heterocycles with



- Arylboronic Acids. *J Am Chem Soc* **2010**, *132* (38), 13194–13196. <https://doi.org/10.1021/ja1066459>.
- (150) Fujiwara, Y.; Domingo, V.; Seiple, I. B.; Gianatassio, R.; Del Bel, M.; Baran, P. S. Practical C–H Functionalization of Quinones with Boronic Acids. *J Am Chem Soc* **2011**, *133* (10), 3292–3295. <https://doi.org/10.1021/ja111152z>.
- (151) Proctor, R. S. J.; Phipps, R. J. Recent Advances in Minisci-Type Reactions. *Angew. Chem., Int. Ed.* **2019**, *58* (39), 13666–13699. <https://doi.org/10.1002/anie.201900977>.
- (152) Baxter, R. D.; Liang, Y.; Hong, X.; Brown, T. A.; Zare, R. N.; Houk, K. N.; Baran, P. S.; Blackmond, D. G. Mechanistic Insights into Two-Phase Radical C–H Arylations. *ACS Cent Sci* **2015**, *1* (8), 456–462. <https://doi.org/10.1021/acscentsci.5b00332>.
- (153) Shi, Y.-F.; Yang, Z.-X.; Ma, S.; Kang, P.-L.; Shang, C.; Hu, P.; Liu, Z.-P. Machine Learning for Chemistry: Basics and Applications. *Engineering* **2023**, *27*, 70–83. <https://doi.org/10.1016/j.eng.2023.04.013>.
- (154) Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.* **1926**, *28* (6), 1049–1070. <https://doi.org/10.1103/PhysRev.28.1049>.
- (155) Born, M.; Oppenheimer, R. Zur Quantentheorie Der Molekeln. *Ann Phys* **1927**, *389* (20), 457–484. <https://doi.org/10.1002/andp.19273892002>.
- (156) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J Chem Phys* **1971**, *54* (2), 724–728. <https://doi.org/10.1063/1.1674902>.
- (157) Slater, J. C. Atomic Shielding Constants. *Physical Review* **1930**, *36* (1), 57–64. <https://doi.org/10.1103/PhysRev.36.57>.
- (158) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829–5835. <https://doi.org/10.1063/1.467146>.
- (159) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *J Chem Phys* **1992**, *97* (4), 2571–2577. <https://doi.org/10.1063/1.463096>.

- (160) Weigend, F.; Furche, F.; Ahlrichs, R. Gaussian Basis Sets of Quadruple Zeta Valence Quality for Atoms H–Kr. *J Chem Phys* **2003**, *119* (24), 12753–12762. <https://doi.org/10.1063/1.1627293>.
- (161) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297–3305. <https://doi.org/10.1039/B508541A>.
- (162) Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24* (1), 89–110. <https://doi.org/DOL: 10.1017/S0305004100011919>.
- (163) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81* (3), 385–390. <https://doi.org/10.1103/PhysRev.81.385>.
- (164) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J Am Chem Soc* **1985**, *107* (13), 3902–3909. <https://doi.org/10.1021/ja00299a024>.
- (165) Pople, J. A.; Beveridge, D. L. *Approximate Molecular Orbital Theory*; McGraw-Hill: New York, 1970.
- (166) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J Mol Model* **2013**, *19* (1), 1–32. <https://doi.org/10.1007/s00894-012-1667-x>.
- (167) Purvis III, G. D.; Bartlett, R. J. A Full Coupled-cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys.* **1982**, *76* (4), 1910–1918. <https://doi.org/10.1063/1.443164>.
- (168) Møller, Chr.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), 618–622. <https://doi.org/10.1103/PhysRev.46.618>.
- (169) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136* (3B), B864–B871. <https://doi.org/10.1103/PhysRev.136.B864>.
- (170) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133–A1138. <https://doi.org/10.1103/PhysRev.140.A1133>.

- (171) Perdew, J. P.; Schmidt, K. Jacob's Ladder of Density Functional Approximations for the Exchange-Correlation Energy. *AIP Conf Proc* **2001**, 577 (1), 1–20. <https://doi.org/10.1063/1.1390175>.
- (172) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys Rev Lett* **1996**, 77 (18), 3865–3868. <https://doi.org/10.1103/PhysRevLett.77.3865>.
- (173) Perdew, J. P. Accurate Density Functional for the Energy: Real-Space Cutoff of the Gradient Expansion for the Exchange Hole. *Phys Rev Lett* **1985**, 55 (16), 1665–1668. <https://doi.org/10.1103/PhysRevLett.55.1665>.
- (174) Becke, A. D. A New Mixing of Hartree–Fock and Local Density-functional Theories. *J. Chem. Phys.* **1993**, 98 (2), 1372–1377. <https://doi.org/10.1063/1.464304>.
- (175) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, 98 (7), 5648–5652. <https://doi.org/10.1063/1.464913>.
- (176) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, 112 (1), 289–320. <https://doi.org/10.1021/cr200107z>.
- (177) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. Assessment of the Performance of the M05–2X and M06–2X Exchange–Correlation Functionals for Noncovalent Interactions in Biomolecules. *J Chem Theory Comput* **2008**, 4 (12), 1996–2000. <https://doi.org/10.1021/ct800308k>.
- (178) Gomes, J. R. B.; Fajín, J. L. C.; Cordeiro, M. N. D. S.; Teixeira, C.; Gomes, P.; Pillai, R. S.; Novell-Leruth, G.; Toda, J.; Jorge, M. Density Functional Treatment of Interactions and Chemical Reactions at Interfaces. In *Density Funct. Theory*; Morin, J., Ed.; New York, 2013; pp 1–58.
- (179) Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J Chem Phys* **1955**, 23 (10), 1833–1840. <https://doi.org/10.1063/1.1740588>.
- (180) Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor Chim Acta* **1977**, 44 (2), 129–138. <https://doi.org/10.1007/BF00549096/METRICS>.

- (181) Lu, T.; Chen, F. Atomic Dipole Moment Corrected Hirshfeld Population Method. *J Theor Comput Chem* **2012**, *11* (01), 163–183. <https://doi.org/10.1142/S0219633612500113>.
- (182) Parr, R. G.; Yang, W. Density Functional Approach to the Frontier-Electron Theory of Chemical Reactivity. *J Am Chem Soc* **1984**, *106* (14), 4049–4050. <https://doi.org/10.1021/ja00326a036>.
- (183) Bultinck, P.; Fias, S.; Van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical Thoughts on Computing Atom Condensed Fukui Functions. *J Chem Phys* **2007**, *127* (3), 034102. <https://doi.org/10.1063/1.2749518>.
- (184) Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Physical Chemistry Chemical Physics* **2015**, *17* (9), 6174–6191. <https://doi.org/10.1039/C5CP00288E>.
- (185) Marenich, A. V; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J Phys Chem B* **2009**, *113* (18), 6378–6396. <https://doi.org/10.1021/jp810292n>.
- (186) HANSCH, C.; MALONEY, P. P.; FUJITA, T.; MUIR, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194* (4824), 178–180. <https://doi.org/10.1038/194178b0>.
- (187) Schneider, A.; Hommel, G.; Blettner, M. Linear Regression Analysis. *Dtsch Arztebl International* **2010**, *107* (44), 776–782.
- (188) *Linear regression in Python (using sklearn and statsmodels)*. <https://www.reneshbedre.com/blog/linear-regression.html> (accessed 2024-10-08).
- (189) Ho, T. K. Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*; 1995; Vol. 1, pp 278–282 vol.1. <https://doi.org/10.1109/ICDAR.1995.598994>.
- (190) Ho, T. K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans Pattern Anal Mach Intell* **1998**, *20* (8), 832–844. <https://doi.org/10.1109/34.709601>.
- (191) Heath, D.; Kasif, S.; Salzberg, S. K-DT: A Multi-Tree Learning Method. In *Proc. of the Second Int. Workshop on Multistrategy Learning*; 1993; pp 138–149.

- (192) Quinlan, J. R. Induction of Decision Trees. *Mach Learn* **1986**, 1 (1), 81–106. <https://doi.org/10.1007/BF00116251>.
- (193) Quinlan, J. R. Simplifying Decision Trees. *Int. J. Man-Mach. Stud.* **1987**, 27 (3), 221–234. [https://doi.org/10.1016/S0020-7373\(87\)80053-6](https://doi.org/10.1016/S0020-7373(87)80053-6).
- (194) Kleinberg, E. M. Stochastic Discrimination. *Ann Math Artif Intell* **1990**, 1 (1), 207–239. <https://doi.org/10.1007/BF01531079>.
- (195) Kleinberg, E. M. An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition. *The Annals of Statistics* **1996**, 24 (6), 2319–2349. <https://doi.org/10.1214/aos/1032181157>.
- (196) Kleinberg, E. M. On the Algorithmic Implementation of Stochastic Discrimination. *IEEE Trans Pattern Anal Mach Intell* **2000**, 22 (5), 473–490. <https://doi.org/10.1109/34.857004>.
- (197) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; KDD '16*; J. Assoc. Comput. Mach.: New York, NY, USA, 2016; pp 785–794. <https://doi.org/10.1145/2939672.2939785>.
- (198) Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, S. Uncertainty-Aware Prediction of Chemical Reaction Yields with Graph Neural Networks. *J Cheminform* **2022**, 14 (1), 2. <https://doi.org/10.1186/s13321-021-00579-z>.
- (199) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences* **1982**, 79 (8), 2554–2558. <https://doi.org/10.1073/pnas.79.8.2554>.
- (200) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J Chem Inf Model* **2024**, 64 (1), 9–17. <https://doi.org/10.1021/acs.jcim.3c01250>.
- (201) Wong, F.; Zheng, E. J.; Valeri, J. A.; Donghia, N. M.; Anahtar, M. N.; Omori, S.; Li, A.; Cubillos-Ruiz, A.; Krishnan, A.; Jin, W.; Manson, A. L.; Friedrichs, J.; Helbig, R.; Hajian, B.; Fiejtek, D. K.; Wagner, F. F.; Soutter, H. H.; Earl, A. M.; Stokes, J. M.; Renner, L. D.; Collins, J. J. Discovery of a Structural Class of Antibiotics with Explainable Deep Learning. *Nature* **2024**, 626 (7997), 177–185. <https://doi.org/10.1038/s41586-023-06887-8>.
- (202) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.;

- Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688-702.e13. <https://doi.org/10.1016/j.cell.2020.01.021>.
- (203) Swanson, K.; Walther, P.; Leitz, J.; Mukherjee, S.; Wu, J. C.; Shivnaraine, R. V.; Zou, J. ADMET-AI: A Machine Learning ADMET Platform for Evaluation of Large-Scale Chemical Libraries. *Bioinformatics* **2024**, *40* (7), btae416. <https://doi.org/10.1093/bioinformatics/btae416>.
- (204) Duan, M.; Shao, Q.; Zhou, Q.; Baran, P. S.; Houk, K. N. Why •CF<sub>2</sub>H Is Nucleophilic but •CF<sub>3</sub> Is Electrophilic in Reactions with Heterocycles. *Nat. Commun.* **2024**, *15* (1), 4630. <https://doi.org/10.1038/s41467-024-48949-z>.
- (205) Nippa, D. F.; Atz, K.; Müller, A. T.; Wolfard, J.; Isert, C.; Binder, M.; Scheidegger, O.; Konrad, D. B.; Grether, U.; Martin, R. E.; Schneider, G. Identifying Opportunities for Late-Stage C-H Alkylation with High-Throughput Experimentation and in Silico Reaction Screening. *Commun Chem* **2023**, *6* (1). <https://doi.org/10.1038/s42004-023-01047-5>.
- (206) Ma, Y.; Liang, J.; Zhao, D.; Chen, Y. L.; Shen, J.; Xiong, B. Condensed Fukui Function Predicts Innate C-H Radical Functionalization Sites on Multi-Nitrogen Containing Fused Arenes. *RSC Adv* **2014**, *4* (33), 17262–17264. <https://doi.org/10.1039/c4ra01853b>.
- (207) King-Smith, E.; Faber, F. A.; Reilly, U.; Sinitskiy, A. V.; Yang, Q.; Liu, B.; Hyek, D.; Lee, A. A. Predictive Minisci Late Stage Functionalization with Transfer Learning. *Nat. Commun.* **2024**, *15* (1), 426. <https://doi.org/10.1038/s41467-023-42145-1>.
- (208) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. AutodE: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions. *Angew. Chem., Int. Ed.* **2021**, *60* (8), 4266–4274. <https://doi.org/10.1002/anie.202011941>.
- (209) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J Chem Theory Comput* **2017**, *13* (11), 5780–5797. <https://doi.org/10.1021/acs.jctc.7b00764>.
- (210) Lee, S.; Ermanis, K.; Goodman, J. M. MolE8: Finding DFT Potential Energy Surface Minima Values from Force-Field Optimised Organic Molecules with

New Machine Learning Representations. *Chem Sci* **2022**, *13* (24), 7204–7214. <https://doi.org/10.1039/d1sc06324c>.

- (211) Wubbels, G. G. Use of the Bell–Evans–Polanyi Principle to Predict Regioselectivity of Nucleophilic Aromatic Photosubstitution Reactions. *Tetrahedron Lett* **2014**, *55* (36), 5066–5069. <https://doi.org/10.1016/j.tetlet.2014.07.042>.
- (212) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratch, 2016. Gaussian 16. 2016.
- (213) Smith, J. M.; Dixon, J. A.; DeGruyter, J. N.; Baran, P. S. Alkyl Sulfinates: Radical Precursors Enabling Drug Discovery. *J Med Chem* **2019**, *62* (5), 2256–2264. <https://doi.org/10.1021/acs.jmedchem.8b01303>.
- (214) Vitaku, E.; Smith, D. T.; Njardarson, J. T. Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved Pharmaceuticals. *J Med Chem* **2014**, *57* (24), 10257–10274. <https://doi.org/10.1021/jm501100b>.
- (215) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; De Jong, W. A. NWChem: A Comprehensive and Scalable Open-Source Solution for Large Scale Molecular Simulations. *Comput Phys Commun* **2010**, *181* (9), 1477–1489. <https://doi.org/10.1016/j.cpc.2010.04.018>.
- (216) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at near Chemical Accuracy with Sub-Second Computational Cost. *Nat. Commun.* **2020**, *11* (1). <https://doi.org/10.1038/s41467-020-16201-z>.
- (217) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor Chem Acc* **2008**, *120* (1–3), 215–241. <https://doi.org/10.1007/s00214-007-0310-x>.
- (218) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J Cheminform* **2011**, *3* (10). <https://doi.org/10.1186/1758-2946-3-33>.

- (219) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J Comput Chem* **1996**, *17* (5–6), 490–519. [https://doi.org/10.1002/\(SICI\)1096-987X\(199604\)17:5/6<490::AID-JCC1>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P).
- (220) Landrum, G.; Tosco, P.; Kelley, B.; Ric; sriniker; Cosgrove, D.; gedec; Vianello, R.; NadineSchneider; Kawashima, E.; N, D.; Jones, G.; Dalke, A.; Cole, B.; Swain, M.; Turk, S.; AlexanderSavelyev; Vaucher, A.; Wójcikowski, M.; Take, I.; Probst, D.; Ujihara, K.; Scalfani, V. F.; godin, guillaume; Pahl, A.; Berenger, F.; JLVarjo; Walker, R.; jasondbiggs; strets123. Rdkit/Rdkit: 2023\_03\_1 (Q1 2023) Release. Zenodo April 2023. <https://doi.org/10.5281/zenodo.7880616>.
- (221) *Daylight Theory Manual*. <https://www.daylight.com/dayhtml/doc/theory/index.html> (accessed 2020-04-04).
- (222) Weser, O.; Hein-Janke, B.; Mata, R. A. Automated Handling of Complex Chemical Structures in Z-Matrix Coordinates—The Chemcoord Library. *J Comput Chem* **2023**, *44* (5), 710–726. <https://doi.org/10.1002/jcc.27029>.
- (223) Stewart, J. J. P. MOPAC2016. Stewart Computational Chemistry: Colorado Springs, CO, USA.
- (224) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratch, 2016. Gaussian 16. 2016.
- (225) Ma, Y.; Liang, J.; Zhao, D.; Chen, Y.-L.; Shen, J.; Xiong, B. Condensed Fukui Function Predicts Innate C–H Radical Functionalization Sites on Multi-Nitrogen Containing Fused Arenes. *RSC Adv.* **2014**, *4* (33), 17262–17264. <https://doi.org/10.1039/C4RA01853B>.
- (226) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maclejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, Di.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res* **2018**, *46* (D1), D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>.
- (227) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv Drug Deliv Rev* **2001**, *46* (1), 3–26. [https://doi.org/10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0).



- (228) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corporation: New York, 1958.
- (229) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel V. and Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer P. and Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, 12, 2825–2830.
- (230) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service. *J Chem Doc* **1965**, 5 (2), 107–113. <https://doi.org/10.1021/c160017a018>.