Using Machine Learning to Optimise Asymmetric Hydrogenation Reactions of Tetra-substituted Olefins

Charlotte Louise Cavens

Hughes Hall

October 2025



UNIVERSITY OF CAMBRIDGE

This thesis is submitted for the degree of Doctor of Philosophy.

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

ABSTRACT

The presence of 1,2-contiguous stereocenters represents an especially valuable motif in drug discovery. However, direct access through the asymmetric hydrogenation (AH) of tetra-substituted olefins (TSC=C) remains a persistent challenge in homogenous catalysis. The high steric hindrance of these substrates, combined with the strong catalyst-substrate specificity of the current systems, makes it challenging to achieve high enantioselectivities for novel substrates. As a result, optimisation campaigns generally demand extensive catalyst screening. This thesis explores an alternative strategy to accelerate the optimisation of these transformations using machine learning (ML).

Chapter 1 outlines the key advancements in the AH of TSC=C, introduces the theoretical basis of applying ML to reaction outcome prediction, and reviews recent studies that employ ML in reaction optimisation.

Chapter 2 describes the construction of a curated literature dataset of AH reactions of TSC=C and evaluates a range of ML models for predicting enantiomeric excess (*ee*). Both regression and classification models were constructed, with tree-based ensembles delivering the strongest performance, achieving absolute errors below 10% *ee* and classification accuracies above 80%.

Chapter 3 applies these models to novel experimental AH reactions of TSC=C, evaluating their ability to perform in new areas of chemical space. While predictive accuracy declined on low-*ee* reactions, the results highlight both the potential and current limitations of the ML models in guiding optimisation beyond the boundaries of existing datasets.

Chapter 4 investigates Bayesian optimisation (BO) as an alternative workflow for experimental optimisation of this transformation. Using Gaussian process regression and multi-objective acquisition functions, the iterative BO campaign efficiently explored reaction space and successfully identified conditions that afforded enantioselectivities of up to 93% *ee*, thereby demonstrating BO as a powerful and generalisable strategy for accelerating discovery in this challenging transformation.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Dr Ruth L. Webster, I have wholeheartedly enjoyed my time in your research group and am sincerely thankful for the opportunity. I am deeply grateful for not only your scientific guidance and support, but also for your encouragement throughout the challenges of this research; your reassurance whenever I felt overwhelmed has always helped me regain perspective.

I extend my appreciation to my supervisor at Bath, Dr Matthew Grayson, for his advice and support, and to Dr Neal J. Fazakerley for his mentorship and guidance, both of which have been invaluable to my development as a researcher.

To all the Webster group, past and present – Dr Leah Webster, Nathan Buxton, Abigail Haliday, Oliver Jarvis, Owen Jones, Dr Adam N. Barret, Dr Thomas G. Linford-Wood and Dr Emily E. Pocock – thank you for creating such a supportive and enjoyable environment. I feel truly fortunate to come away from my PhD with so many lifelong friends.

To all members of the Grayson group, past and present, I am indebted for the knowledge and insight you have shared. A special thank you to Dr Sam Espley, Dr Elliot H. E. Farrar and Toby Lewis-Atwell, without whose expertise this research would not have been possible.

I would also like to thank all the technical support staff at the Universities of Cambridge and Bath for their invaluable assistance throughout my research.

Dr Thomas M. Hood and Dr Samantha Y-S. Lau, you are my twin pillars, and I thank you for all you have taught me. I have always looked up to you both and will continue to do so, in my career and beyond.

Dr Mirela A. Johnson, both your brilliance as a scientist and your kindness as a friend have been an inspiration to me; thank you for being my best friend.

Irene V-S and Ella J. Marshall, my girls, thank you for always being there for me, through every high and low, your friendship means the world.

Finally, I would like to thank my family for their support: my brothers, who are no doubt tired of hearing about this work, and above all my mum, my greatest inspiration, whose unwavering love and support has carried me through this journey.

CONTENTS

1 Introduction 12

1.1 SIGNIFICANCE OF RESEARCH 13

1.2 ASYMMETRIC HYDROGENATION REACTIONS OF TETRA-SUBSTITUTED OLEFINS 16

1.2.1 Unfunctionalised Tetra-substituted Olefins 17

1.2.2 Functionalised Tetra-substituted Olefins 23

1.2.3 Mechanistic Investigations of Tetra-substituted Olefins 29

1.3 MACHINE LEARNING THEORY 34

1.3.1 Machine Learning Overview 34

1.3.2 Dataset Generation 36

1.3.3 Feature Generation 39

1.3.4 Dataset Splitting and Model Training 42

1.3.5 Model Evaluation 44

1.3.6 Machine Learning Algorithms 46

1.3.7 Linear Regression Algorithms 47

1.3.8 Decision Tree Algorithms 49

1.3.9 Support Vector Machines 51

1.4 MACHINE LEARNING FOR REACTION OUTCOME PREDICTION 53

1.4.1 Machine Learning for *ee* Prediction in AH Reactions of Olefins 55

1.5 PROJECT AIMS 57

2 Using Machine Learning to Predict Enantiomeric Excess of a Literature Dataset of Asymmetric Hydrogenations of Tetra-substituted Olefins 58

2.1 INTRODUCTION 59

2.1.1 Chapter Aims 59
2.1.2 Computational Theory 59
2.2 METHODS 62
2.2.1 Dataset Generation (i) 62
2.2.2 Computational Methods (ii) 63
2.2.3 Feature Generation (iii) 63
2.2.4 Machine Learning Methods (iv) 66
2.2.5 Machine Learning Evaluation (v) 67
2.3 RESULTS AND DISCUSSION 69
2.3.1 Dataset Analysis 69
2.3.2 Regression Machine Learning Models 77
2.3.3 Classification Machine Learning Models 86
2.3.4 Synthetic Minority Over-sampling of Classification Models 94
2.4 CONCLUSIONS 98
3 Evaluating Machine Learning Models on Novel Asymmetric Hydrogenation Reactions of Tetra-substituted Olefins 100
3.1 INTRODUCTION 101
3.1.1 Chapter Aims 103
3.2 METHODS 104
3.3 RESULTS AND DISCUSSION 106
3.3.1 Prediction of Reactions of a Novel Substrate 106
3.3.2 Predictions of Literature Substrates 107
3.4 CONCLUSIONS AND FUTURE WORK 115
3.5 EXPERIMENTAL 116

3.5.1 General Considerations 116
3.5.2 General Procedure for Asymmetric Hydrogenation Reactions 116
3.5.3 General Procedure for Racemic Hydrogenation Reactions 117
3.5.4 Characterisation of Hydrogenated Product 2,3-Diphenyl-indan-1-one 117
3.5.5 Preparation of Substrates 118
3.5.6 Characterisation of Hydrogenated Products 122
3.5.7 Synthesis of Iridium Complex 125

4 Applying Bayesian Optimisation to Asymmetric Hydrogenation Reactions of Tetra-substituted Olefins 126

4.1 INTRODUCTION 127

4.1.1 Acknowledgment of Contributing Researchers 127
4.1.2 Bayesian Optimisation 127
4.1.3 Gaussian Process Regression 128
4.1.4 Chemical Representations for Bayesian Optimisation 129
4.1.5 Bayesian Optimisation for Reaction Optimisation 131
4.1.6 Multi-objective Bayesian Optimisation 133
4.1.7 Asymmetric Hydrogenation of Tetra-substituted Enamides 136
4.1.8 Chapter Aims 142

4.2 METHODS 143

4.2.1 Reaction Space 144
4.2.2 Model Features 146
4.2.3 Initial Reactions 146
4.2.4 Machine Learning Methods 146
4.2.5 Conversion Tolerance 147

4.2.6 Batch Selection 149

4.3 RESULTS AND DISCUSSION 150

4.3.1 Principal Component Analysis 150

4.3.2 Initial Set of Reactions 151

4.3.3 Optimisation One 157

4.3.4 Optimisation Two 162

4.3.5 Optimisation Three 166

4.3.6 Optimisation Four 170

4.3.7 Optimisation Five 174

4.3.8 Summary of Bayesian Optimisation Performance 177

4.4 CONCLUSIONS AND FUTURE WORK 185

4.5 EXPERIMENTAL 186

4.5.1 General Considerations 186

4.5.2 Synthesis of *N*-(2-Methylcyclohept-1-en-1-yl)acetamide 187

4.5.3 Synthesis of *N*-(-2-Methylcycloheptyl)acetamide 189

4.5.4 General Procedure for the Asymmetric Hydrogenation Reactions of *N*-(-2-Methylcycloheptyl)acetamide 190

5 Conclusion 191

6 References 193

7 Appendix 213

7.1 CHAPTER 2 APPENDIX 213

1.1 CHAPTER 3 APPENDIX 241

1.2 CHAPTER 4 APPENDIX 252

ABBREVIATIONS

acac	Acetylacetone
ADASYN	Adaptive synthetic sampling
AF	Acquisition function
AH	Asymmetric hydrogenation
AI	Artificial intelligence
BAr^F_4	bis(triflourmethyl)phenyl borate
BIPI	Boehringer-Ingelheim phosphinoimidazoline
BO	Bayesian optimisation
CNN	Convolutional neural networks
COD	Cyclooctadiene
CREST	Conformer-rotamer ensemble sampling tool
CV	Cross-validation
DCM	Dichloromethane
de-F	Defluorinated
DFT	Density functional theory
DFTB	Density functional tight binding
DIPAMP	1,2-Bis[(2-methoxyphenyl)(phenyl)phosphino]ethane
DNN	Deep neural network
DOE	Design of experiments
DOI	Digital Object Identifier
E-states	Electronic states
EBTHI	Ethylenebis(tetrahydroindenyl)
ECFPs	Extended-connectivity fingerprints
ee	Enantiomeric excess
Et-DuPhos	1,2-Bis(2,5-diethylphospholano)benzene
ETC	Extra-trees classifier
ETR	Extra-trees regressor
fEI	Expected Improvement
GBC	Gradient boosting classifier
GBR	Gradient boosting regressor
GC	Genetic cross-over
GFN	Geometry, frequency, noncovalent
GPR	Gaussian Process Regressor

HBF_4	Tetrafluoroboric acid
HOMO	Highest occupied molecular orbital
HTE	High-throughput experimentation
InChI	International chemical identifier
JAK	Janus kinase
KDE	Kernel density estimation
KHMDS	potassium bis(trimethylsilyl)amide
L-DOPA	L-3,4-Dihydroxyphenylalanine
LASSO	Least absolute shrinkage and selection operator
LLR	Lasso Linear Regressor
LogP	Partition coefficient
LR	Linear regression
LUMO	Lowest unoccupied molecular orbital
MAC	methyl (*Z*)-α-acetamidocinnamate
MAE	Mean absolute error
MeOBIPOP	6,6′-Dimethoxy-2,2′-bis(diphenylphosphino)-1,1′-biphenyl
MFF	Multiple fingerprint feature
ML	Machine learning
MLR	Multiple linear regression
MM	Molecular mechanics
MTD	Metadynamics
$NaBAr^F{}_4$	Tetrakis[3,5-bis(trifluoromethyl)phenyl]borate
NBD	Norbornadiene
NN	Neural network
OHE	One-hot encoding
PBV	Percent Buried Volume
PC1	Principal component 1
PC2	Principal component 2
PCA	Principal component analysis
Pd/C	Palladium on carbon
PES	Potential energy surface
Ph-BPE	1,2-Bis[-2,5-diphenylphospholano]ethane
q-EHVI	q-Expected Hypervolume Improvement
QC	Quantum chemical
QM	Quantum mechanical

QSPR	Quantitative structure–property relationships
R^2	Coefficient of determination
RBF	Radial Basis Function
RDKit	Rational discovery kit
RDS	Rate determining step
RFC	Random forest classifier
RFR	Random forest regressor
RR	Ridge Regressor
RMSD	Root mean squared deviation
RMSE	Root mean squared error
rt	Room temperature
SASA	Solvent accessible surface area
SD	Standard deviation
SFC	Supercritical fluid chromatography
sklearn	Scikit learn
SMILES	Simplified Molecular Input Line Entry System
SMOTE	Synthetic minority over-sampling Technique
SVC	Support vector classification
SVM	Support vector machines
SVR	Support vector regressor
TCFP	Trichickenfootphos
TEA	Triethylamine
TFA	Trifluoroacetic acid
THF	Tetrahydrofuran
TPSA	Total polar surface area
TS	Transition state
TS-EMO	Thompson sampling efficient multi-objective optimisations
TSC=C	Tetra-substituted olefin
UFF	Universal force field
xTB	Extended tight-binding
$\Delta G^{\ddagger}$	Activation free energy

1 Introduction

1.1 SIGNIFICANCE OF RESEARCH

The growing demands of modern drug discovery, particularly the pursuit of candidates with enhanced potency and selectivity, have intensified the need for innovative synthetic methodologies.[1, 2] Among the most pressing challenges is the development of efficient routes to compounds containing 1,2-contiguous stereocenters.[3, 4] Asymmetric hydrogenation (AH) of tetra-substituted olefins (TSC=C) offers a direct and powerful strategy to access this important motif,[5-7] yet it remains one of the most challenging transformations in asymmetric catalysis.[8]

The significance of this challenge is underscored by the fact that many pharmaceutically relevant and natural product compounds contain 1,2-contiguous stereocenters as defining features of their bioactive scaffolds (**Figure 1.1**).[4-7, 9-13] The critical importance of enantioselective synthesis stems from the profound influence of chirality on biological function.[14-16] More than one-third of marketed pharmaceuticals are chiral molecules and enantiomeric pairs often display drastically different pharmacological profiles,[17] reflecting the inherent chirality of proteins and enzymes that interact with them.[18] In many cases, one enantiomer is therapeutically active while the other is inactive or causes adverse effects. Well-known examples include thalidomide and ethambutol, where the enantiomers differ dramatically in their biological outcomes.[19, 20] Although racemic synthesis followed by enantiomeric resolution is possible, such approaches are inefficient, atom-wasteful, and intrinsically low-yielding.[18] By contrast, catalytic methods that deliver a single enantiomer directly represent a more sustainable and scalable strategy. Thus, the AH of TSC=C is an especially attractive goal for pharmaceutical synthesis, due to its formation of two-contiguous stereocenters in a single step.

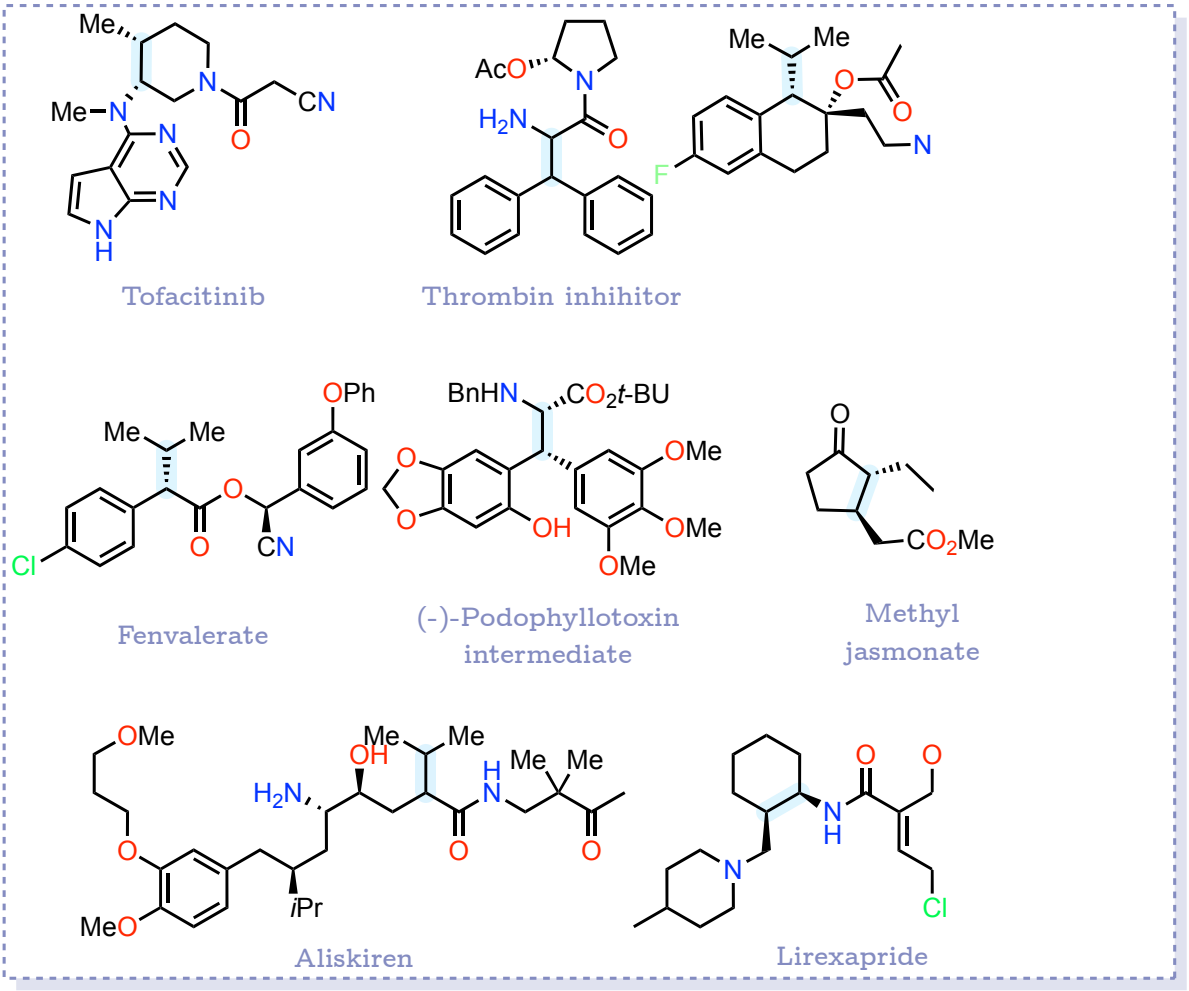
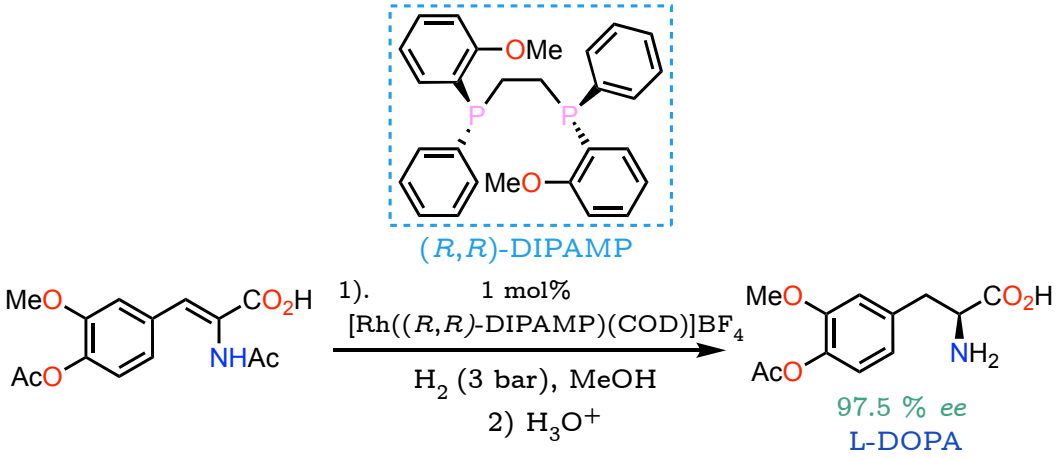


Figure 1.1: Chemical structures of representative pharmaceutically- and biologically-relevant compounds.

The significance of AH reactions was firmly established when the 2001 Nobel Prize in Chemistry was awarded to Noyori and Knowles for their pioneering contributions in this field.[21-23] Notably, Knowles employed a rhodium complex with the chiral bisphosphine ligand 1,2-bis[(2-methoxyphenyl)(phenyl)phosphino]ethane (DIPAMP) to achieve the enantioselective synthesis of L-3,4-dihydroxyphenylalanine (L-DOPA), a key treatment for Parkinson's disease (**Scheme 1.1**).[24] Since this landmark achievement, thousands of examples of AH applied to di- and tri-substituted olefins have been reported.[25] In contrast, AH of tetra-substituted olefins has proven far more difficult and consequently remains relatively underexplored.



Scheme 1.1: Synthesis of L-DOPA developed by Knowles and co-workers. AH followed by an acid hydrolysis obtained L-DOPA.

Traditionally, optimisation of this challenging reaction has relied on exhaustive trial-and-error experimentation with model substrates, a process that is both time-consuming and resource-intensive. This strategy also depends heavily on expert chemical intuition and a deep mechanistic understanding of the reaction. In recent years, however, machine learning (ML) has emerged as a powerful tool to accelerate reaction development.[26] By identifying patterns in existing datasets that correlate with successful outcomes, ML models can guide the design of future experiments more efficiently.[27] Crucially, predictive models applied at the early stages of screening in an optimisation campaign when only limited data is available can significantly accelerate discovery of chiral catalysts and conditions that lead to successful reactions.[28]

Integrating such data-driven approaches into the experimental workflow of AH reactions of TSC=C has the potential to reduce failed reactions, shorten optimisation timelines, and conserve valuable resources. This can be achieved through the prediction of enantiomeric excess (*ee*) and identifying conditions that deliver both high *ee* and yield.

1.2 ASYMMETRIC HYDROGENATION REACTIONS OF TETRA-SUBSTITUTED OLEFINS

The lack of reports on AH of TSC=C, in contrast to their di- and tri-substituted counterparts,[25] is primarily due to the increased steric hindrance of this motif. Typically, the carbon that coordinates to the metal catalyst is the less sterically encumbered carbon, which is rarely the prochiral olefin in di- and tri-substituted substrates. Consequently, it is more difficult both to differentiate between the carbons of tetra-substituted substrates and achieve effective coordination.[3] As a result, the development of chiral catalysts for such substrates has posed significant synthetic challenges. Moreover, these catalysts often lack generality. Currently, no single catalyst is compatible across all classes of TSC=C. In fact, existing catalyst systems tend to be highly substrate-specific, where even minor structural modifications can lead to a substantial loss in selectivity.[29]

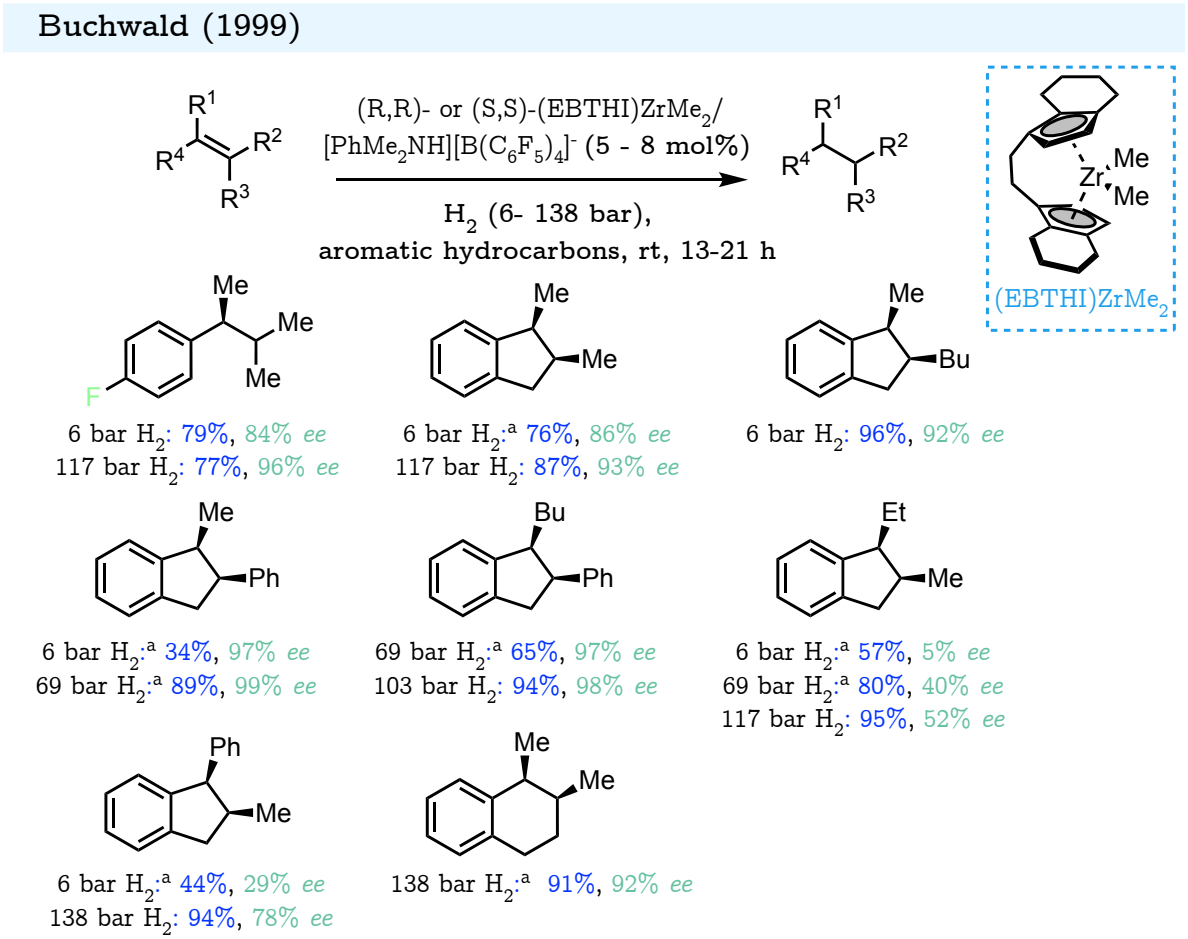
Many successful catalytic systems have been developed, employing a variety of ligand scaffolds in combination with mainly precious metals such as rhodium, iridium, and ruthenium. However, a reliance on these metals highlights a persistent problem in homogenous catalysis due to their scarcity and high cost.[30]

In the successful reports of AH of TSC=C, substrates are often classified as either 'functionalised' or 'unfunctionalised', depending on the presence or absence of a coordinating group directly conjugated to the carbon-carbon double bond.[31] Functionalised substrates, which contain such groups, can coordinate directly to the metal centre, thereby orienting the molecule around the catalyst. These substrates account for the majority of reported reactions in this area.[32] Importantly, the term unfunctionalised does not refer exclusively to olefins with purely aliphatic substituents, but more broadly to substrates lacking a specific coordinating group. For instance, while substrates containing aromatic rings are generally categorised as unfunctionalised, these substituents can engage in interactions with the metal centre.[31]

This section outlines key advancements in the literature on AH reactions of tetra-substituted functionalised and unfunctionalised olefins and highlights general trends and mechanistic insights into these transformations.

1.2.1 Unfunctionalised Tetra-substituted Olefins

The primary challenge associated with unfunctionalised substrates, is due to this lack of an additional 'handle' for the metal catalyst to coordinate. However, in 1999 Buchwald and coworkers exploited the high electrophilicity of cationic zirconocene catalysts for the AH of unfunctionalised indenes (**Scheme 1.2**).[33] The active catalyst was formed *in-situ* under hydrogen pressure from (*S*,*S*)- or (*R*,*R*)-ethylenebis(tetrahydroindenyl)dimethylzirconium ((*S*,*S*)−(EBTHI)ZrMe$_2$]) and dimethylanilinium perfluorotetraphenylborate.



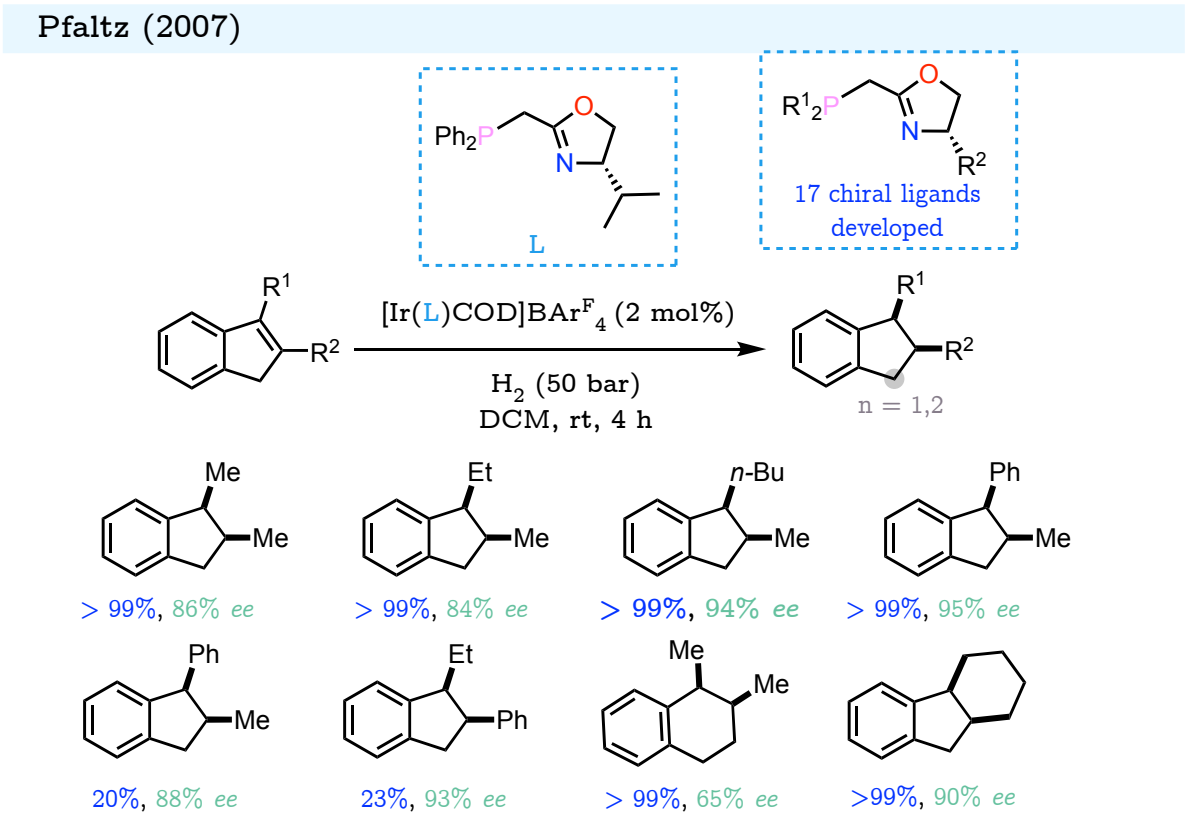
Scheme 1.2: The AH of unfunctionalised substrates reported by Buchwald. The reactions were run at rt for 13-21 h, with 8 and [a]5 mol% catalyst loading in aromatic hydrocarbon solvents under 6 – 138 bar H_2. Conversions are reported in blue and were determined by gas chromatography and ^{1}H NMR. *ee* is reported in green.

This study represented the first successful AH of unfunctionalised TSC=C substrates. However, the protocol was not widely adopted, likely due to the unpredictable enantioselectivities observed and the sensitivity of the catalyst system. The mechanism

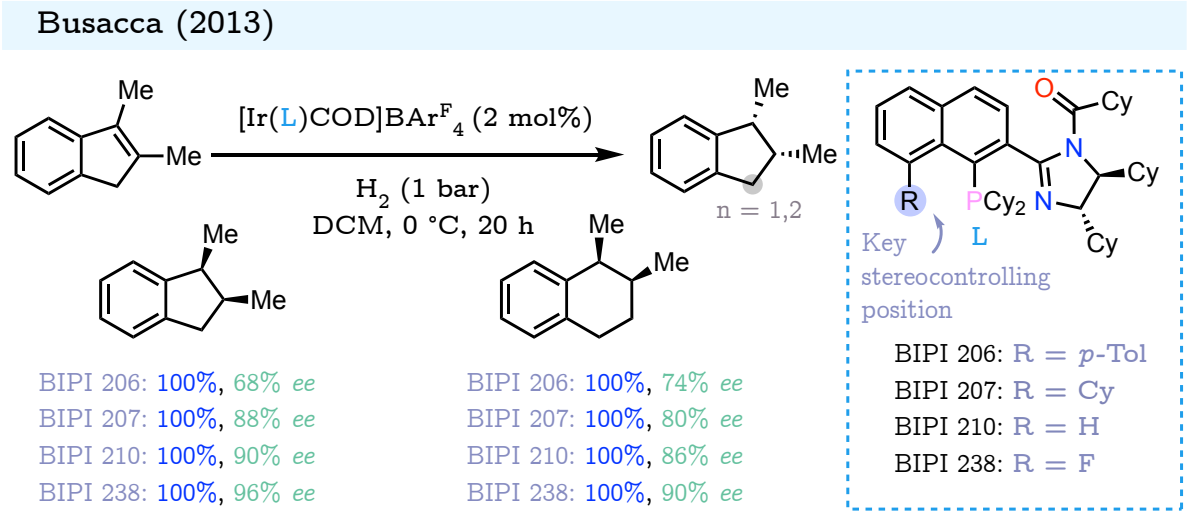
was highly complex and substrate-dependent, with enantioselectivity strongly influenced by hydrogen pressure. Notably, higher pressures were required to achieve both increased conversion and improved enantioselectivity on certain substrates, particularly those bearing phenyl substituents.

The next successful report of AH of unfunctionalised TSC=C was in 2007 when Pfaltz reported the first use of chiral bidentate P,N ligands for this transformation (**Scheme 1.3**).[34] After screening a library of 17 chiral ligands containing the phophanyl oxazoline motif, the ligand shown in **Scheme 1.3** (first reported by Sprinz and Helmchen),[35] gave the optimum result and was employed with bis(1,5-cyclooctadiene)iridium (1) tetrakis[3,5-bis(trifluoromethyl)phenyl]borate ([Ir(COD)][BAr^F_4]). This opened a new class of ligands for this reaction. The substrate scope presented shows this catalyst obtained higher enantioselectivities at a lower hydrogen pressure than that achieved using the chiral zirconocene catalyst developed by Buchwald. However, this system also had limitations reporting a narrow substrate scope with only moderate enantioselectivities achieved for dihydronapthalene (65% *ee*).

In the search for another effective chiral ligand for this transformation, Busacca reported a ligand optimisation study on the use of Boehringer-Ingelheim phosphinoimidazoline (BIPI) ligands in [Ir(COD)][BAr^F_4] complexes.[36]A series of BIPI ligands were synthesised, and those bearing naphthyl and *N*-acyl substituents exhibited superior performance. Notably, the *peri* position (C8) of the naphthyl core was identified as a key stereocontrolling site, and a range of ligands with different substituents at this position were evaluated (**Scheme 1.4**). Although difficulties in ligand synthesis prevented preparation of a larger library, the fluorine substituent proved optimal, obtaining 96% *ee* at low hydrogen pressures for the AH of 2,3-dimethylindene.

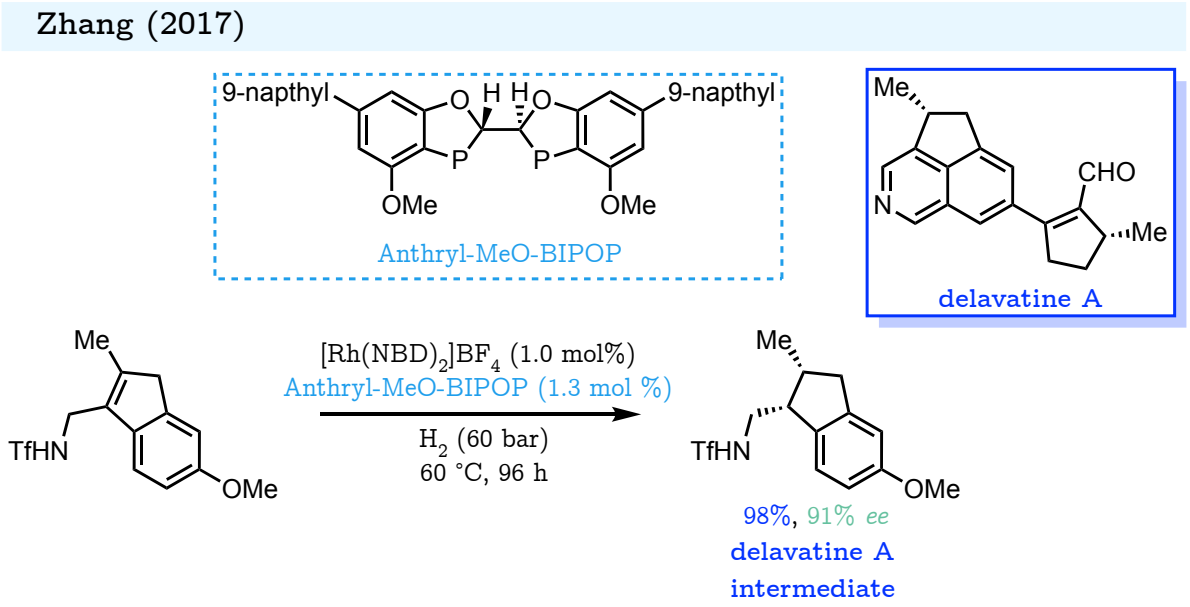


Scheme 1.3: The AH of unfunctionalised substrates reported by Pfaltz. The reactions were run at rt for 4 h, with 2 mol% catalyst loading in dichloromethane (DCM) under 50 bar H_2. Conversions are reported in blue and *ee* is reported in green.



Scheme 1.4: The optimum results of the ligand optimisation study reported by Bussaca. The reactions were run at 0 °C for 20 h, with 2 mol% catalyst loading in DCM under 1 bar H_2. Conversions are reported in blue and *ee* is reported in green.

The Zhang group highlighted the importance of AH of TSC=C substrates in drug discovery, through their synthesis of delevatine A,[37] a compound with anticancer properties. A series of bisphosphine chiral ligands were screened against a similar motif as the delavatine A intermediate. In contrast to earlier reports employing the [Ir(COD)][BAr^F_4] fragment, [Rh(bisphosphine)]BF_4 complexes were generated in-situ from bis(norbornadiene)rhodium(I) tetrafluoroborate ([$Rh(NBD)_2$]BF_4).[38] The screen identified 6,6′-dimethoxy-2,2′-bis(diphenylphosphino)-1,1′-biphenyl (MeOBIPOP) as a promising ligand, reported previously used by the Tang group for the AH of functionalised enamides.[39] However, a slightly modified anthryl-substituted MeOBIPOP, featuring deep chiral pockets, proved optimal for the gram-scale synthesis of a key intermediate of delavatine A (**Scheme 1.5**).

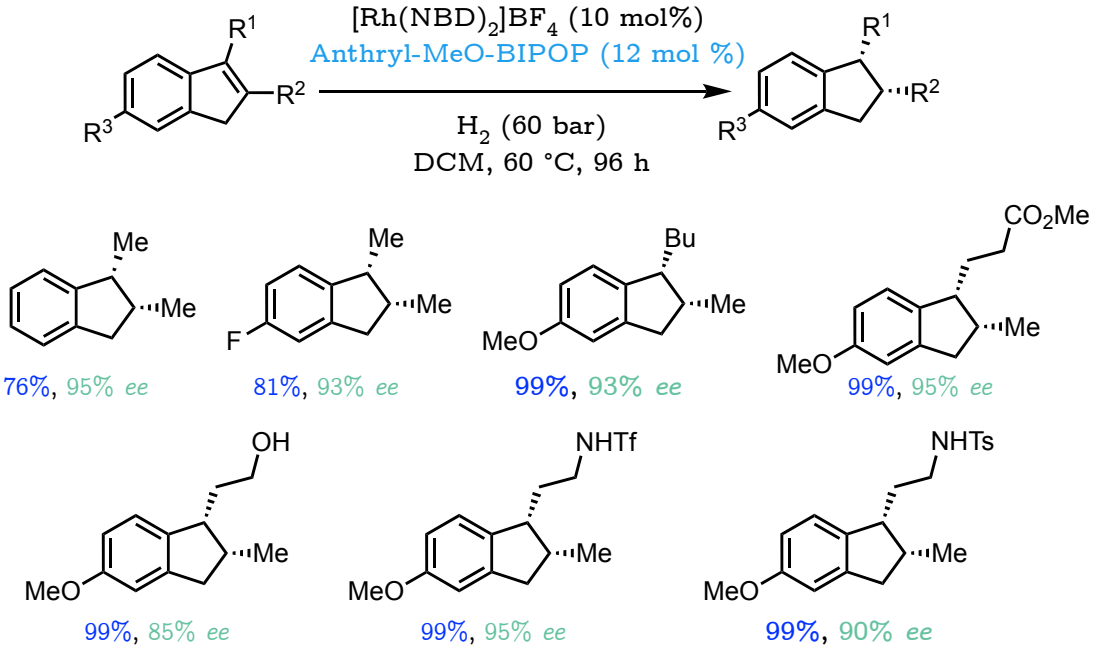


Scheme 1.5: AH of delavatine A intermediate. The reaction was conducted under 60 bar H_2 at 60 °C for 96 hrs. Rh(Anthryl-MeO-BIPOP)]BF_4 was generated in-situ from 10 mol% [$Rh(NBD)_2$]BF_4 and 12 mol% anthryl-MeOBIPOP. The resultant yield of the reaction is reported in blue and *ee* in green.

The catalytic system was further applied to a broader range of unfunctionalised substrates (**Scheme 1.6**). A variety of indene-type substrates were hydrogenated with good enantioselectivities, and the system was found to tolerate both electron-withdrawing and electron-donating substituents on the phenyl ring. Moreover, the addition of more polar functional groups, such as sulfonamides and alcohols, did not significantly disrupt catalytic performance. This study demonstrated a more diverse

substrate scope than previously reported. However, the reactions required relatively harsher conditions, including elevated temperature (60 °C), higher hydrogen pressure (60 bar), increased catalyst loading (10 mol%), and extended reaction times (up to 96 h).

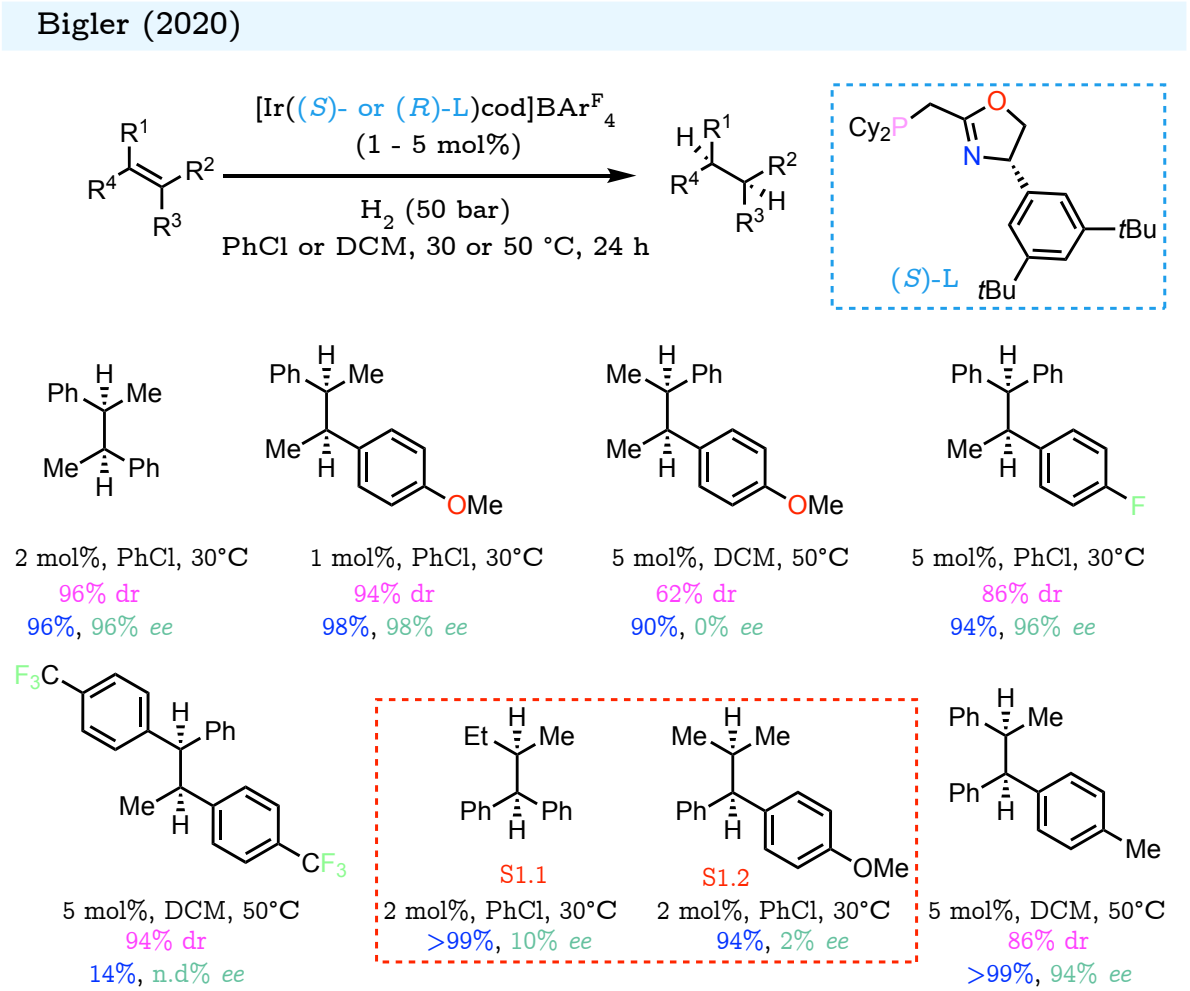
Zhang (2017)



Scheme 1.6: Representative examples of the substrate scope repored by Zhang. The reaction was conducted under 60 bar H_2 at 60 °C for 96 hrs in DCM. Rh(Anthryl-MeO-BIPOP)]BF_4 was generated in-situ from 10 mol% [Rh(NBD)$_2$]BF_4 and 12 mol% anthryl-MeOBIPOP. The resultant yield of the reaction is reported in blue and *ee* in green.

The enhanced functional group tolerance reported in this study represented a significant advance in the AH of TSC=C substrates. A main limitation of the reports up to this point is that the majority of substrates were cyclic, commonly featuring an indene-type scaffold. In 2020, however, a study was published which targeted the AH of acyclic tetra-substituted olefins.[8] A screen of 34 diverse ligand structures coordinated to [Ir(COD)][BAr^F_4] complexes was conducted for the AH of (*E*)-2,3-diphenylbut-2-ene. Phosphanyl oxazoline ligands were found to outperform both bisphosphine and ferrocene-containing ligands, motivating the synthesis and secondary screening of 15 structural variants of this class of ligands. This effort identified the ligand shown in **Scheme 1.7** as optimal, notably containing the same substructure previously reported by Pfaltz in 2007.[34]

A comprehensive substrate scope was subsequently reported, with representative examples shown in **Scheme 1.7**. A broad range of substrates were tolerated, and high enantioselectivities were achieved, including for challenging alkyl–triaryl olefins. In contrast, substrates bearing two trifluoromethyl substituents were not compatible with the system. Furthermore, reduced enantioselectivities were observed for substrates **S1.1** and **S1.2**. In which, the faces of the olefin are only differentiated by the size of the alkyl substituent for **S1.1**, and the addition of a methoxy group on the phenyl ring for **S1.2**.



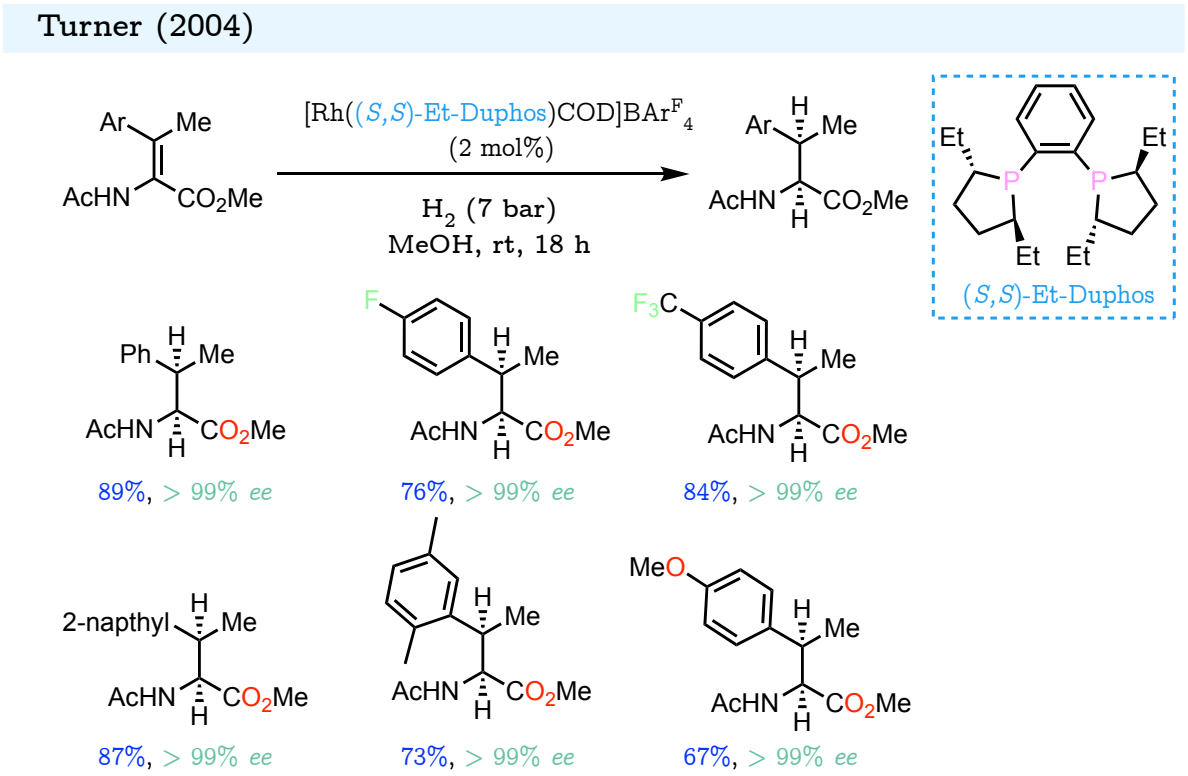
Scheme 1.7: Representative examples from the scope of acyclic unfunctionalised TSC=C reported by Bigler. The reactions were run at 30 °C or 50 °C for 24 h, with 1 -5 mol% catalyst loading in DCM or chlorobenzene (PhCl) under 50 bar H_2. Conversions are reported in blue, diastereomeric ratio as % in pink and *ee* is reported in green.

1.2.2 Functionalised Tetra-substituted Olefins

Functionalised olefins constitute the majority of reported AH of TSC=C substrates.[40] This is likely attributable to the enhanced reactivity conferred by coordinating groups toward the metal catalyst compared with unfunctionalised substrates. In addition, demand for the hydrogenated products of this class is higher, as they include biologically-relevant motifs such as chiral amino acids and carboxylic acids.[3]

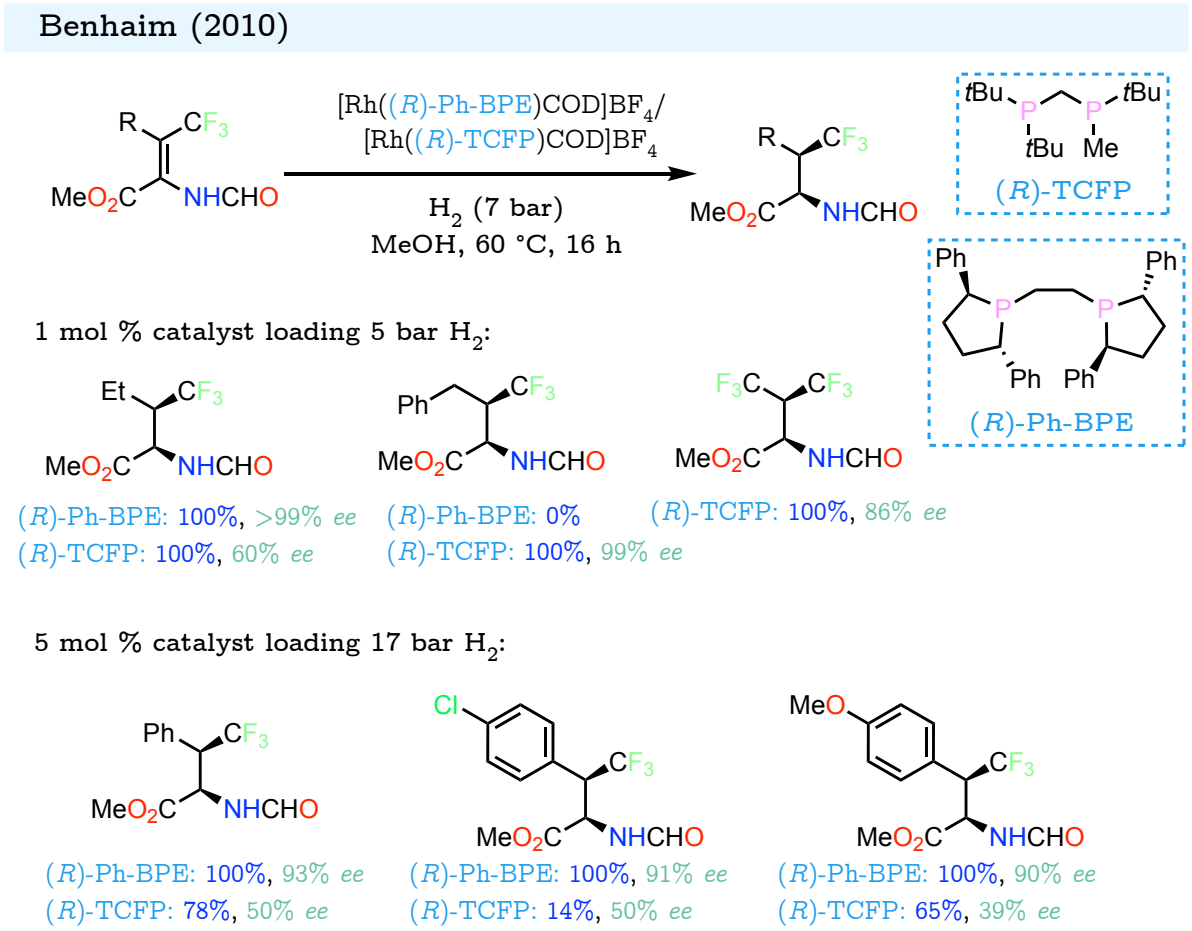
A pivotal breakthrough in this area was reported by Burk in 1995 with the AH of both cyclic and acyclic tetra-substituted enamides using cationic bisphosphine–rhodium catalysts.[41] This work paved the way for the development of other successful catalytic systems for tetra-substituted enamides, which are discussed in greater detail in Chapter 4.

In 2004, Turner reported the AH of a series of (*Z*)-didehydroamino acid derivatives, which, upon deprotection, yielded β-branched α-amino acids.[42] This transformation underscored its synthetic significance, as β-branched α-amino acids are key building blocks for the preparation of modified peptides with enzyme-inhibitory activity. Optimal conditions employed (*S*,*S*)-1,2-bis(2,5-diethylphospholano)benzene ((*S*,*S*)-Et-Duphos) in rhodium(I) (1,5-cyclooctadiene) tetrafluoridoborate ($Rh(COD)BF_4$) complexes. Although the reported substrate scope was relatively narrow (**Scheme 1.8**), the system delivered high enantioselectivities at low hydrogen pressures and was compatible with both electron-donating and electron-withdrawing substituents on the phenyl ring.



Scheme 1.8: The AH scope reported by Turner. The reactions were run at rt for 18 h, with 2 mol% catalyst loading in methanol (MeOH) under 7 bar H_2. Yields are reported in blue and *ee* is reported in green.

In 2010, Benhaim and co-workers expanded on this work by developing the AH of trifluoromethyl-substituted substrates to access β-trifluoromethyl α-amino acids (**Scheme 1.9**).[29] A comprehensive ligand screen covering multiple ligand families identified the commercially-available (*R*)-1,2-bis[-2,5-diphenylphospholano]ethane ((*R*)-Ph-BPE) as particularly effective. This ligand, along with (*R*)-Trichickenfootphos ((*R*)-TCFP), originally developed by Hoge and coworkers,[43] was employed in [Rh(COD)]BF_4 complexes. The substrate scope revealed a high degree of catalyst specificity, whereby small structural modifications led to significant losses in reactivity and enantioselectivity. For example, replacing an ethyl substituent with a benzyl group eliminated reactivity with (*R*)-Ph-BPE, whereas (*R*)-TCFP provided full conversion with excellent enantioselectivity. Conversely, substrates bearing aromatic substituents required increased catalyst loading and hydrogen pressure to proceed. However, under these conditions, (*R*)-TCFP performed more poorly reflected in poorer conversions and enantioselectivities of 50% *ee* or lower. In these cases, (*R*)-Ph-BPE proved to be the optimal ligand.



Scheme 1.9: The AH scope reported by Benhaim. The reactions were run at 60 °C for 16 h, with 1 -5 mol% catalyst loading in MeOH under 5 – 17 bar H_2. Yields are reported in blue and *ee* is reported in green.

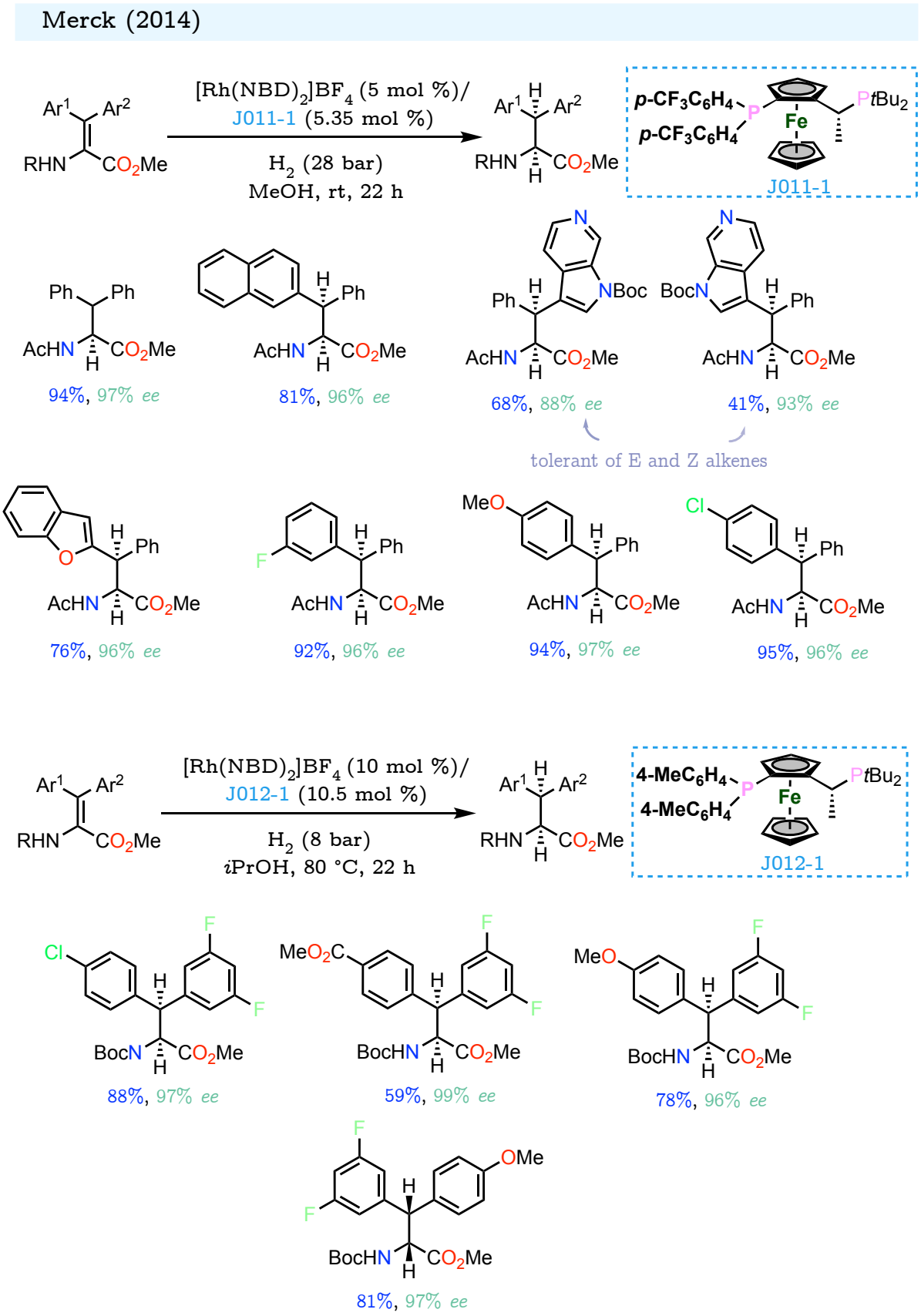
The synthesis of β,β-diaryl-α-amines bearing non-identical β-aryl substituents presents a considerable challenge, particularly in retaining defined *E*/*Z* geometries during both substrate preparation and subsequent AH. Two teams at Merck successfully developed such substrates in high yield and achieved efficient AH.[44] Whereas ruthenium- and iridium-based catalysts exhibited poor reactivity, rhodium catalysis—specifically using bis(norbornadiene)rhodium(I) tetrafluoridoborate ($[Rh(NBD)_2]BF_4$) as the precatalyst and generating the chiral catalyst *in-situ*—delivered promising results. In an extensive optimisation campaign screening 150 chiral ligands, C1-symmetric planar chiral ferrocene ligands emerged as highly reactive and selective.

Using the optimal ligands J011-1 and J012-1, a range of *N*-methoxycarbonyl β,β-diaryl dehydroamino acids were hydrogenated successfully (**Scheme 1.10**). High enantioselectivities were observed regardless of whether the phenyl substituents

contained electron-donating or electron-withdrawing groups. Notably, the catalytic system was effective for both *E*- and *Z*-alkenes. In addition, a broad array of aryl and heteroaryl substituents was tolerated, as well as the less reactive Boc-protected amines, for which J012-1 was identified as the optimal ligand.

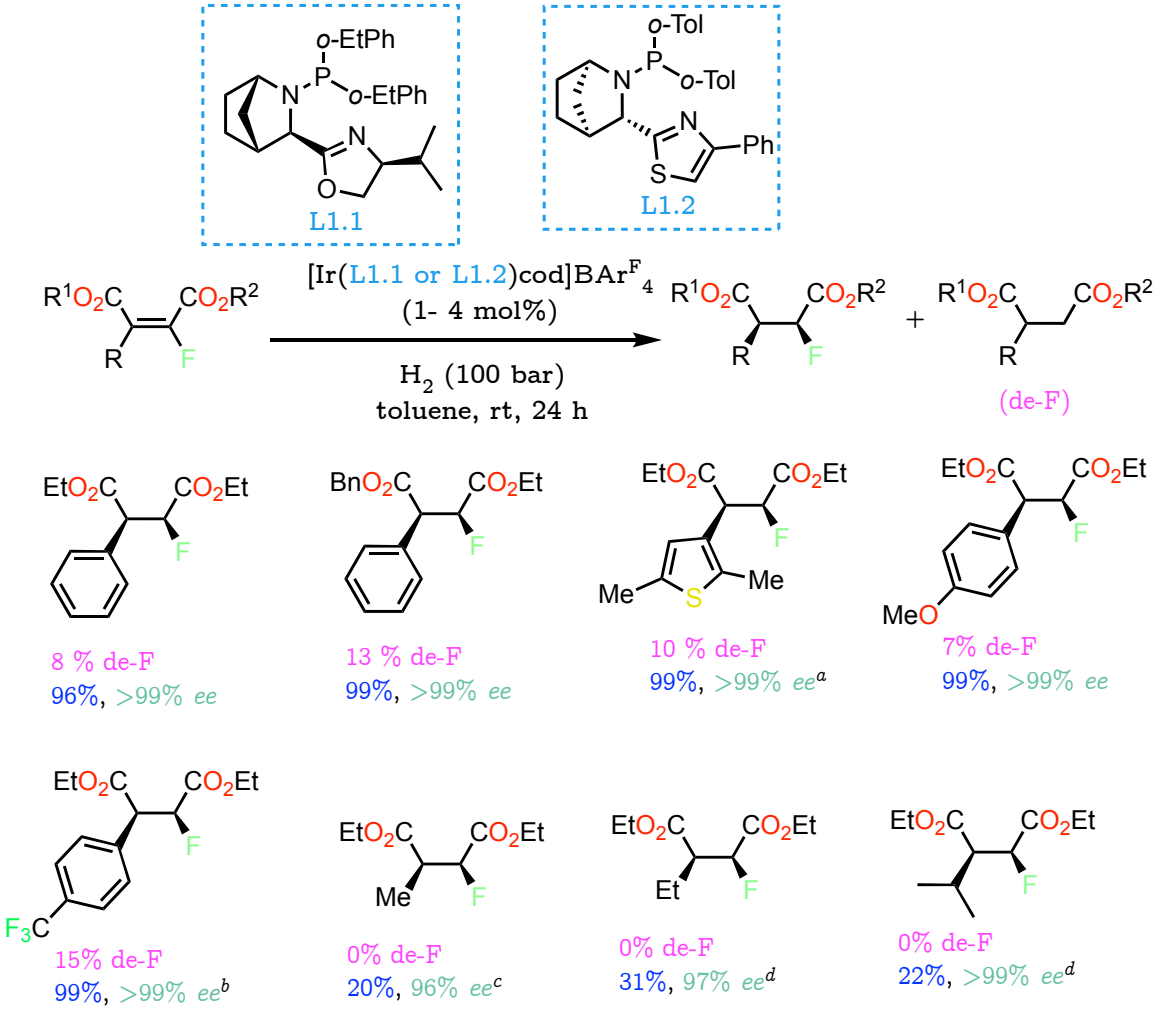
Few examples in this field have included synthetically useful handles, such as halides, silanes, or boranes, which could serve as sites for onward chemical transformations.[3] In 2018, Andersson and co-workers reported the AH of unactivated vinyl fluorides, building on the work by Pfaltz using chiral ligands with [Ir(COD)][BAr^F_4] complexes. A ligand screen of bicyclic Ir P,N complexes containing either thiazole or oxazoline rings identified a new ligand, **L1.1** (**Scheme 1.11**).[45] The catalytic system was evaluated on a series of (*E*)-2-fluoroacrylate substrates, delivering excellent enantioselectivities for several examples.

Unsurprisingly, the greatest challenge with this substrate class was formation of the defluorinated (de-F) byproduct, arising from the susceptibility of C–F bonds to metal-mediated activation and cleavage under hydrogenation conditions. The de-F product was observed in the majority of cases, underscoring the current incompatibility of incorporating such synthetic handles into AH methodologies. Nevertheless, a diverse substrate scope was demonstrated in which, both electron-withdrawing and electron-donating substituents were well tolerated, although increased catalyst loading (from 2 to 4 mol%) was required for substrates bearing dimethylthienyl and CF_3 groups. In contrast, substrates with aliphatic substituents showed significantly reduced conversions and *ee*, necessitating a switch from **L1.1** to the alternative ligand **L1.2**.



Scheme 1.10: Representative examples of the AH scope reported by two teams at Merck. The reactions were run at rt for 22 h, with 5 mol% $[Rh(NBD)_2]BF_4$ and 5.35 mol% J011-1 in methanol under 28 bar H_2, or 80 °C for 22 h, with 5 mol% $[Rh(NBD)_2]BF_4$ and 5.35 mol% J012-1 in isopropanol (iPrOH) under 8 bar H_2.

Andersson (2017)



Scheme 1.11: Representative examples of the AH of tetra-substituted vinyl flourides reported by Andersson. The reactions were run at rt for 24 h, with 1-4 mol% catalyst loading in toluene under 100 bar H_2. [a] 2 mol% catalyst loading, [b] 4 mol% catalyst loading, [c] 2 mol% catalyst loading of L1.2, [d] 4 mol% catalyst loading of L1.2 Conversions are reported in blue and include both hydrogenated product and deflourinated byproduct (de-F), *ee* is reported in green.

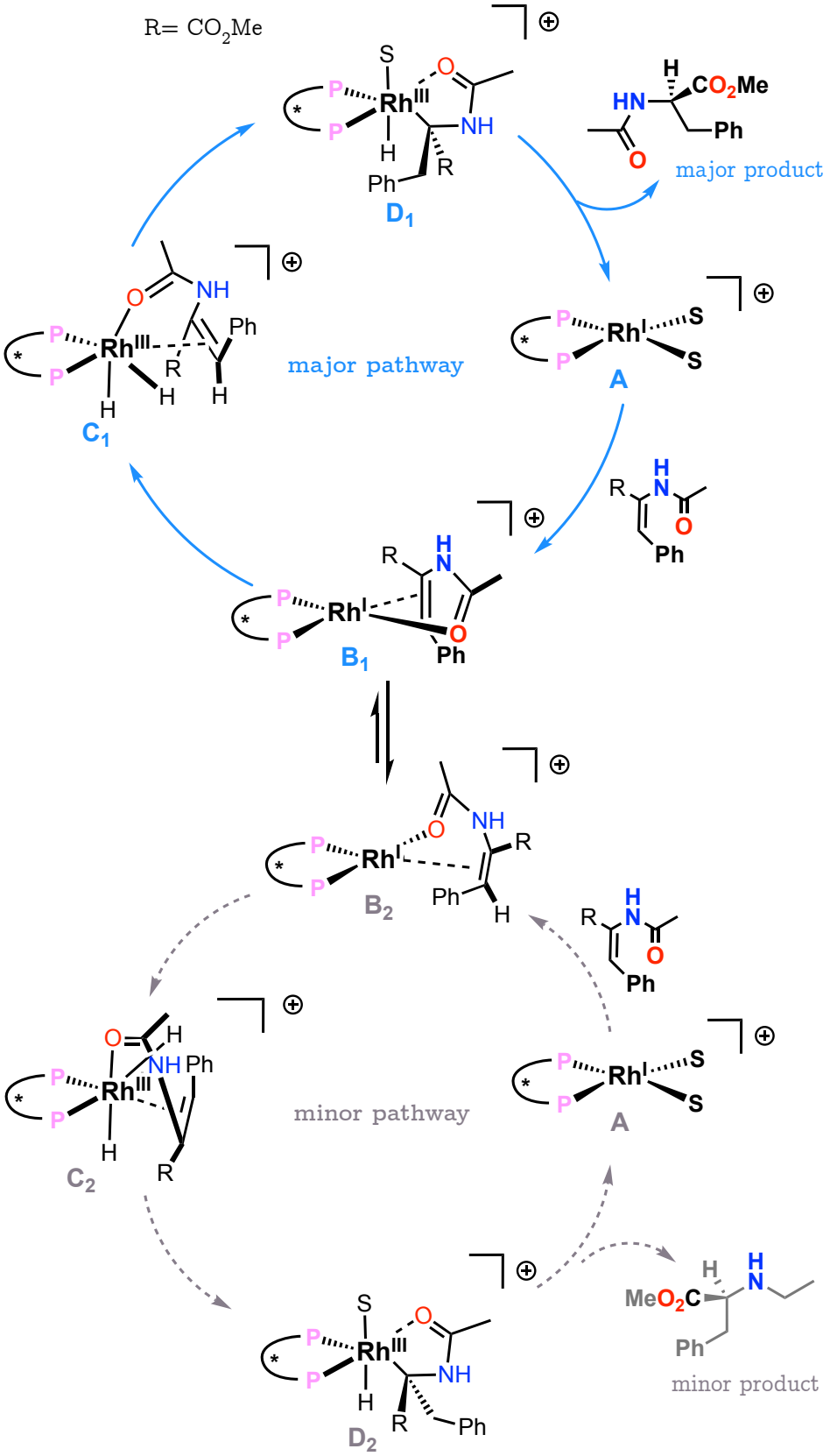
1.2.3 Mechanistic Investigations of Tetra-substituted Olefins

The AH of TSC=C remains one of the most demanding challenges in homogeneous catalysis. However, notable successes have been achieved and demonstrate overarching themes from the literature. Chiral catalysts in this field are highly substrate-specific, where even minor structural modifications to the substrate often result in dramatic changes in catalytic performance. This lack of generality means that optimisation campaigns typically begin with the screening of model substrates against a large ligand library, followed by the systematic evaluation of substrate scope, an exhaustive and time-consuming approach.

Additionally, some approximate trends can be derived between functionalised and unfunctionalised substrates. Since Pfaltz's seminal pioneering report, the AH of unfunctionalised olefins has been dominated by Ir–P,N complexes, most commonly with oxazoline scaffolds employed with [Ir(COD)][BAr^F_4] complexes. In contrast, functionalised substrates continue to be more extensively studied with bisphosphine ligands employed in rhodium catalysts often generated in-situ.

The mechanism of AH of TSC=C remains comparatively underexplored in the literature. In many cases, it is assumed that the mechanism proceeds through pathways analogous to those established for tri-substituted alkenes, despite the additional steric and electronic complexity of tetra-substituted substrates. Among transition-metal catalysts, rhodium was the most extensively investigated. However, mechanistic proposals for Rh-catalysed AH of TSC=C substrates are still largely assumed from studies on tri-substituted olefins.

A commonly reported mechanistic example of rhodium AH employs a chiral bisphosphine–Rh(I) complex, specifically with (*S*,*S*)-1,2-bis[(2-methoxyphenyl)(phenyl)phosphino]ethane ((*S*,*S*)-DIPAMP) for the hydrogenation of the tri-substituted olefin, dehydroamino acid methyl (Z)-α-acetamidocinnamate (MAC, **Scheme 1.12**).



Scheme 1.12: Mechanistic example of Rh-catalysed AH of MAC. The major pathway is shown in blue and the minor pathway is shown in grey.

The favoured mechanism shown in **Scheme 1.12** is known to be the unsaturated mechanism.[46, 47] In this mechanism, the 16-electron square planar complex, **A** coordinates to either the *si* or *re* face of the olefin resulting in the two square planar complexes $\mathbf{B_1}$ and $\mathbf{B_2}$, respectively. Notably, for this case of a functionalised substrate, the olefin coordinates in a bidentate manner, with both the directing group and the olefin unit coordinated to the metal centre. Additionally, the alkene coordination is a dynamic equilibrium, and in this instance a greater proportion of $\mathbf{B_2}$ was observed. Following alkene coordination, oxidative addition of dihydrogen results in the formation of an 18-electron octahedral complex ($\mathbf{C_1}$ and $\mathbf{C_2}$). This is the rate determining step (RDS) of the reaction.[48] Subsequently, migratory insertion of the alkene into an Rh-H bond forms $\mathbf{D_1}$ and $\mathbf{D_2}$. Finally, reductive C-H elimination results in the formation of the two enantiomers. Complexes $\mathbf{B_1}$ and $\mathbf{B_2}$ are diastereomers which ultimately govern enantioslectivity leading to the two enantiomeric products.

In the example above, the major enantiomer from $\mathbf{D_1}$ is selectively formed over that from $\mathbf{D_2}$. As discussed, coordination of the olefin resulting in $\mathbf{B_2}$ is favoured relative to $\mathbf{B_1}$ (*si*-face binding). However due to the Curtin-Hammett principle,[49] the rate of oxidative addition (the RDS) for $\mathbf{B_1}$ is significantly higher than $\mathbf{B_2}$ due to the steric clashes that result in complexes $\mathbf{C_1}$ and $\mathbf{C_2}$, resulting in the selective formation of the major enantiomer from $\mathbf{D_1}$.

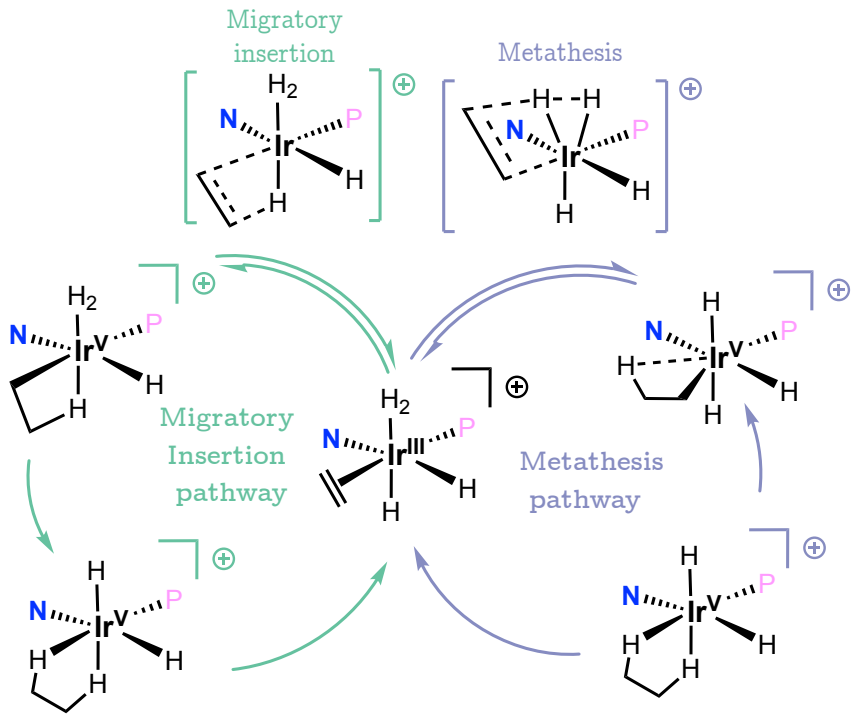
Although the unsaturated pathway discussed above is the most common, a dihydride pathway has also been reported specifically for electron-rich phosphine ligands,[50, 51] in which the oxidative addition of dihydrogen to **A** occurs before substrate coordination. The remaining mechanistic pathway is analogous to the unsaturated pathway reported.

Similarly, variations in mechanism are observed for different metal centres. While the AH reactions facilitated by rhodium operate through the mentioned Rh(I) and Rh(III) centres, ruthenium catalysts employ Ru(II) and Ru(IV) oxidation states in an analogous mechanism.

In contrast to the mechanistic understanding of rhodium-catalysed AH, which is largely derived from studies on tri-substituted olefins, recent years have seen a

growing number of reports examining iridium-catalysed AH of TSC=C.[8, 52] The reports have focused on systems employing chiral *P,N*-oxazoline ligands, reflecting their remarkable success in the hydrogenation of unfunctionalised olefins. Mechanistic investigations by Andersson, Bigler, and Burgess have shown that the Ir-catalyzsd process most likely operates through Ir(III)/Ir(V) hydride species over the analogous Ir(I)/Ir(III) seen in Rh-based systems.[53-56] Notably, the operative pathway has been reported as a pressure-dependent system, with the Ir(I)/Ir(III) operating at low hydrogen pressure.[57, 58]

A representative Ir(III)/Ir(V) mechanism is illustrated in **Scheme 1.13**.[59] The alkene substrate typically coordinates to the dihydride-dihydrogen complex, generating $[(N,P^*)Ir^{III}(H)_2H_2)(alkene)]^+$.[3, 60, 61] Subsequently, the enantiodetermining hydrogen transfer to the coordinated olefin can proceed by two possible pathways: (i) migratory insertion of a hydride from the metal centre into the C=C bond, or (ii) σ-bond metathesis between dihydrogen and the olefin. These processes generate an Ir(V)–alkyl–hydride intermediate (or related species), which is stabilized by the steric and electronic environment of the ligand. Subsequent reductive elimination releases the hydrogenated product, with H_2 activation regenerating the active catalyst. In 2014, Pfaltz and co-workers succeeded in characterising Ir–dihydride complexes including $[(N,P^*)Ir^{III}(H)_2H_2)(alkene)]^+$,[57] lending experimental support to this mechanism. Their work also concluded that hydrogen was required to convert the alkene bound to the catalyst to the hydrogenated product. Additional computational studies from other groups suggest that the migratory insertion pathway is generally preferred to the σ-bond metathesis.[55, 56, 62]



Scheme 1.13: Mechanistic example of Ir-catalysed AH of an unfunctionalised olefin. The migratory insertion pathway is shown in green and the metathesis pathway is shown in purple.

1.3 MACHINE LEARNING THEORY

In principle, ML can be applied to predict the experimental *ee* of AH reactions of TSC=C, thereby reducing the resources and time typically required for trial-and-error experimentation in optimisation campaigns.

In this section, the general workflow and theory of applying ML to chemical problems of this nature is outlined.

1.3.1 Machine Learning Overview

ML is a rapidly developing subfield of artificial intelligence (AI) that focuses on developing computational models capable of identifying patterns within data. Over the past decade, ML has found applications across all domains of science, [63-67] and its use within chemistry has been increasingly widespread.[68-83] These applications include reaction outcome prediction,[26, 68, 84, 85], prediction of DFT outputs,[86-89] materials property prediction,[76] and in spectroscopy and analytical chemistry.[90-92]

ML methods are broadly categorised into supervised and unsupervised learning (**Figure 1.2**).[93] In supervised learning, algorithms are trained using labelled data where both inputs and corresponding outputs are known. The model learns an objective function by optimising its internal parameters so that it can predict the target outputs for previously 'unseen' data. In contrast, unsupervised learning is trained only on input data without labelled outputs. Instead, it identifies hidden structures or relationships within datasets, such as in clustering or dimensionality reduction tasks. For example, in chemistry, unsupervised methods can be used to cluster chemical structures based on similar structural motifs.[94] However, the majority of ML applications in chemical research fall under the domain of supervised learning.[93]

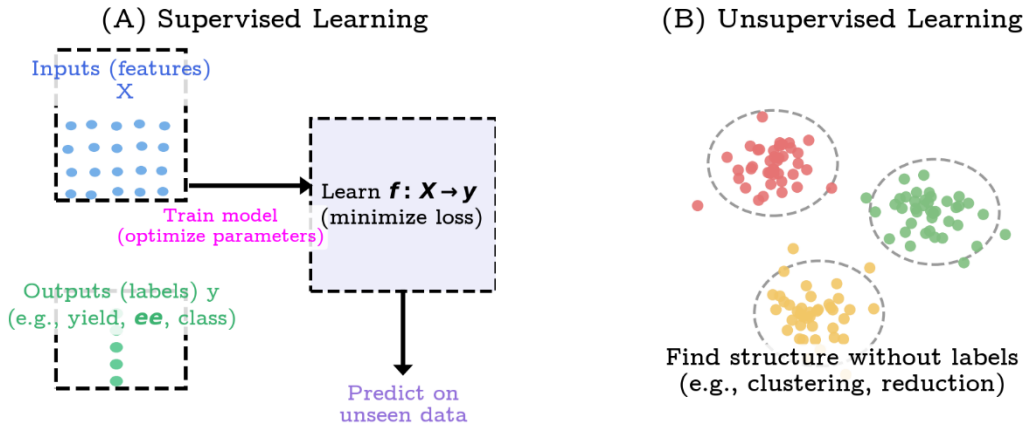


Figure 1.2: Overview of ML approaches. A) Supervised learning: models are trained on paired inputs and output labels to learn a mapping that can then be used to make predictions on unseen data. B) Unsupervised learning: models are trained only on input features, with no labelled outputs, to uncover patterns or trends in the data.

Supervised learning problems are typically further divided into classification and regression tasks (**Figure 1.3**). In classification models outputs belong to discrete categories (e.g., atom type or reaction class), whereas in regression models, outputs are continuous values (e.g. reaction yield or *ee*).[95] The input data, termed 'features', are critical for model performance. In chemical ML, careful featurisation or encoding of molecular structures is particularly important as it strongly influences the accuracy and generalisability of the model.[96] Generalisability, in the context of ML, refers to the ability of the model to make reliable predictions on 'unseen' data not included during training.

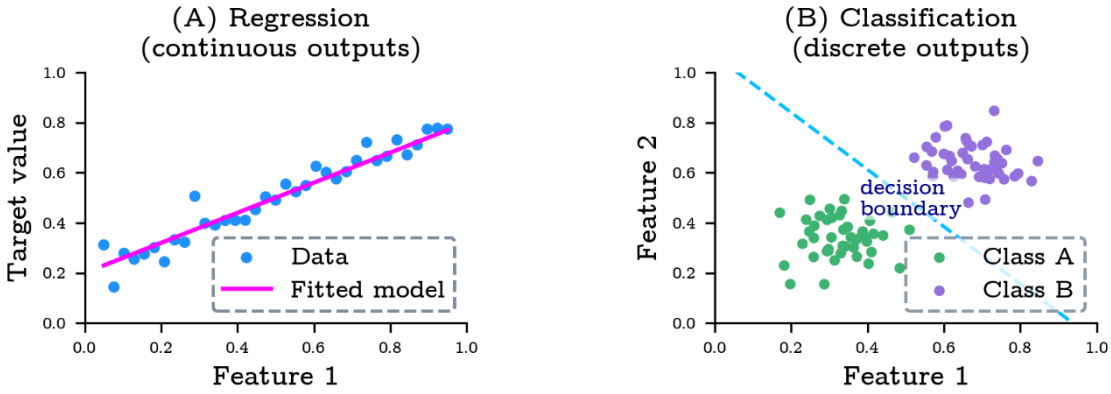


Figure 1.3: Example of supervised learning tasks. A) Regression: a continuous-valued output predicted from input data (blue), the magenta line shows a fitted linear model. B) Classification: data points are assigned to a category (Class A in green and Class B in purple) based on their input features. The dashed blue line indicates the linear decision boundary between two classes.

In addition to feature representation, both the size and diversity of the training dataset significantly impact predictive performance. These considerations will be discussed in more detail in Sections **1.3.2** and **1.3.3**, with specific emphasis on how they apply to predicting experimental results.

The workflow for building a predictive supervised ML model begins with the collection of an appropriate dataset. Next, machine-interpretable features are generated, an ML algorithm is selected, and the dataset is divided into training and test sets. The model is then trained on the input features to learn the target objectives, and its performance is evaluated by assessing how accurately it predicts target outputs for previously unseen data.

1.3.2 Dataset Generation

The generation of a suitable dataset is typically the first step in applying ML methods to chemical problems. The size and quality of the dataset have a profound impact on model performance.[97] However, the amount of data required depends on both the complexity of the problem domain and the class of model selected.

A logical starting point for dataset generation is the use of data already available in the published literature. Although the process of collecting such data can be time-consuming, it remains a practical and cost-effective strategy. However, chemical reactions are inherently high-dimensional, involving numerous parameters such as solvent, additives, temperature, catalyst identity, catalyst loading, and reaction time. Consequently, the chemical reaction space is orders of magnitude larger than the conceptual chemical space defined only by the molecular structures of the reactants and products.[98] Available reaction data therefore cover only a narrow subset of conditions and are often reported inconsistently, leading to incomplete datasets.[99]

Owing to these limitations and the effort required to manually curate datasets from the literature, there has been a strong push in recent years toward the creation of publicly accessible reaction databases.[95] These resources have become increasingly prevalent, driven both by the growing availability of published reaction data and the high demand for accurate reaction outcome prediction. Developing predictive models

from such datasets is therefore highly desirable for both academic and industrial applications.[76]

In addition to the considerable time required to curate datasets from the literature, another major challenge lies in the biases inherent in published experimental data. These biases have been broadly categorised into three types (**Figure 1.4**).[98] First, experimental noise, which arises from unavoidable errors and variability in experimental execution. Second, selection bias, driven by the tendency of researchers to explore reaction conditions that have previously proven successful, thereby neglecting unexplored conditions that might also yield favourable outcomes. Third, reporting bias, since literature sources are typically skewed toward positive results, with 'failed' or low-yielding experiments rarely disclosed.[100] This reporting bias can result in datasets that are imbalanced and therefore not fully representative of the problem domain.



Figure 1.4: Sources of bias in reaction data used for machine learning. A) Experimental noise, arising from errors during data collection. B) Selection bias, reflecting the tendency of researchers to preferentially explore reaction conditions that are already known or expected to work well, leaving much of the reaction space unexplored. C) Reporting bias, caused by the underreporting of low-yielding or 'failed' experiments, leading to datasets skewed to positive outcomes.

A dataset is considered imbalanced when the target outputs are not equally represented. For example, an experimental dataset may predominantly contain reactions with high yields or high *ee* (majority class), while unsuccessful or low-performing reactions (minority class) are rarely reported in publications and may only be accessible through archived laboratory notebooks.[100, 101] Such imbalance can hinder

model generalisability. One straightforward strategy to mitigate this issue is to resample the dataset, either by over-sampling the minority class or under-sampling the majority class (naïve sampling). However, such methods can be inefficient, specifically for the production of ML methods for experimental predictions.[102]

A more sophisticated strategy for addressing imbalanced datasets is the generation of synthetic data. The synthetic minority over-sampling technique (SMOTE) creates new data points for the minority class by interpolation.[103] Specifically, a data point from the minority class is randomly selected, and a line is drawn between it and one of its nearest neighbours in feature space. A synthetic sample is then generated at a random point along this line. Adaptive synthetic sampling (ADASYN) is a modification of SMOTE in which the number of synthetic samples generated is guided by the local density distribution: more synthetic points are created in sparse regions of feature space that are more difficult for the model to learn.[104] A representation of these methods is shown in **Figure 1.5**.

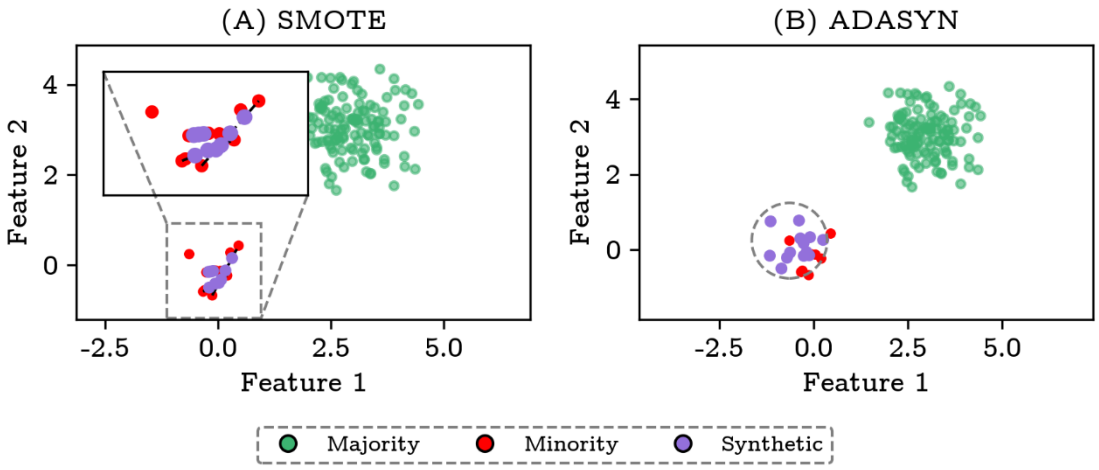


Figure 1.5: Illustration of synthetic oversampling strategies for imbalanced datasets. A) The synthetic minority oversampling technique (SMOTE) B) Adaptive synthetic sampling (ADASYN).

Although synthetic data generation can improve model training, its effectiveness is inherently limited, and a larger volume of real experimental data remains preferable. In this context, high-throughput experimentation (HTE) is a particularly powerful approach. By enabling many parallel experiments to be conducted simultaneously, HTE rapidly produces large datasets that better capture the underlying distribution

of outcomes. Many of these datasets have been used for the generation of successful predictive ML models.[68, 105] Moreover, the combination of HTE and ML offers additional value through their self-reinforcing nature: ML methods can guide the design of HTE campaigns, and the resulting data can be iteratively fed back into the ML model, establishing a closed feedback loop that accelerates discovery.[106]

Despite significant recent progress in automated HTE, its availability remains limited, particularly within academia due to instrumentation costs and infrastructure demands.[107] Furthermore, current platforms are often not compatible with harsh conditions such as high temperatures and pressures, and the datasets they generate can still suffer from considerable experimental noise.

Given the incomplete and noisy nature of experimental datasets, it is perhaps unsurprising that the majority of ML applications in chemistry to date have focused on computational datasets, where data can be systematically constructed and maintained with high consistency.[108] Such datasets are often used to predict physicochemical properties from input features derived from quantum mechanical (QM) calculations, typically density functional theory (DFT).[86] ML models trained on such datasets have been remarkably successful, as evidenced by hundreds of publications reporting accurate predictions.[109] However, the integration of ML into the experimental workflows of synthetic chemists remains rare.[108, 110] Developing predictive models that can directly leverage experimental data to deliver practically relevant insights is therefore a critical next step.

1.3.3 Feature Generation

While data collection provides the foundation for machine learning, the subsequent representation of this data as features is equally important, as it directly influences model performance. Well-designed features capture the essential aspects of the problem domain and enable the model to establish a reliable mapping between inputs and outputs.[111] In many cases, features can be transformed into a lower-dimensional representation to reduce complexity while preserving the most relevant information, although such transformations may also lead to information loss. In chemistry, features

typically consist of machine-interpretable encodings of molecular structures and/or their physicochemical properties (**Figure 1.6**).[112]

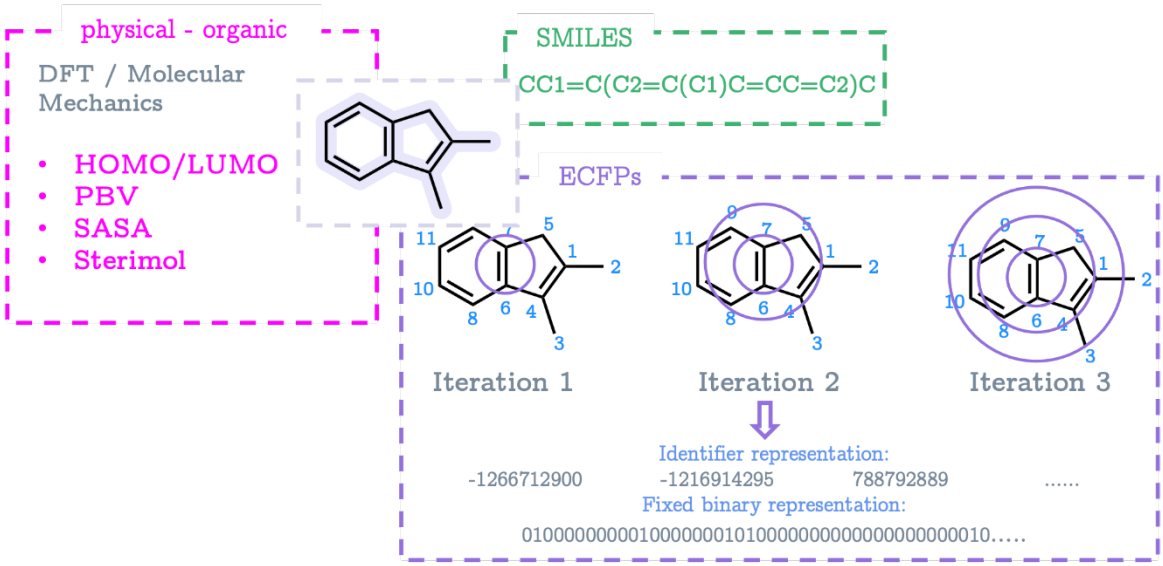


Figure 1.6: Illustrations of example features used for chemical structures in supervised ML tasks.

There are various methods used in chemistry to encode molecular structures. However, retaining the structural information or key properties that would lead to accurate predictions of the target output is a key challenge.[113, 114] Often in computational chemistry structures are represented by their three-dimensional coordinates or Cartesian coordinates. However, raw Cartesian coordinates are generally unsuitable as features for ML, since effective molecular representations must be invariant to rigid transformations, such as translation and rotation.[115] Alternative representations include molecular fingerprints,[116, 117] International Chemical Identifier (InChI) keys,[118] molecular graphs,[119] and simplified molecular input line entry strings (SMILES).[120, 121]

In SMILES, molecules are encoded as character strings representing atoms and bonds, and this representation has been successfully applied in tasks such as virtual screening and lead discovery. A limitation of SMILES is that they are non-canonical, meaning that multiple valid SMILES strings can represent the same molecular structure. This redundancy can lead to complications in structure searching and may introduce data leakage in ML tasks.[122] To address this, algorithms have been developed to canonicalize SMILES and ensure that a unique string corresponds to each structure.[123-125] Unfortunately, because SMILES were originally designed to represent

organic molecules with covalent bonding, they lack a formalism for describing coordination or dative bonds, such as those found in metal–ligand complexes.[126] By contrast, this limitation does not apply to molecular fingerprints which, due to their ease of generation and richer chemical information content, are often the preferred choice for ML applications.[112]

Molecular topological fingerprints represent chemical structures as bit vectors. Among these, the extended-connectivity fingerprints (ECFPs) are particularly popular.[117] ECFPs are circular topological fingerprints generated through a variant of the Morgan algorithm.[127] The procedure begins by assigning every non-hydrogen atom an integer identifier, which is iteratively updated based on the identifiers of neighbouring atoms. With each iteration, progressively larger atom neighbourhoods are captured. These identifiers are then hashed into integers, duplicates are removed, and the resulting set is represented as bits in a virtual bit string. For use in ML, where features must have fixed length, the virtual string is folded into a fixed-size bit vector.

Beyond structural encodings, steric and electronic or physical-organic descriptors also provide interpretable features for supervised ML models. Simple descriptors include counts of functional groups or chemical motifs (e.g., hydrogen-bond donors, hydrogen-bond acceptors, aromatic rings, sp^3 carbons). More complex descriptors are derived from QC or from molecular mechanics (MM) optimised structures. Examples include frontier orbital energies (highest occupies molecular orbital, HOMO, lowest unoccupied molecular orbital, LUMO), partition coefficient (logP),[128] atomic charges, electronegativity,[129] Sterimol parameters,[130] percentage buried volume (%Vbur), electrotopological states (E-states), and solvent-accessible surface area (SASA).

Since physical-organic descriptors are often expressed in different units and scales, preprocessing is essential. ML models generally achieve higher performance when input features are scaled to a comparable range.[131] Two standard approaches are normalisation and standardisation: in normalisation, each feature value is scaled to lie within the range 0–1, whereas in standardisation, values are centred by subtracting the mean and scaled by the standard deviation (SD). Both approaches facilitate

pattern recognition by the learning algorithm and reduce bias toward features with larger absolute magnitudes.

QM derived descriptors can capture more complex chemical information than basic structural encoding or simple descriptors but often require prior mechanistic knowledge of the reaction class being modelled in order to extract useful features. Moreover, descriptors derived from QM calculations, such as those based on DFT, can be computationally expensive and time-consuming to generate.[132] Consequently, when appropriate, the use of less complex representations such as molecular fingerprints may provide a more practical alternative to QM-derived features.[133]

1.3.4 Dataset Splitting and Model Training

Following the collection of data and the generation of input features, the dataset must be split into a training set and a test set. Conventionally in supervised ML, this is achieved through a train–test split, in which the data is separated into two subsets. This is most commonly done randomly, but other methods, such as scaffold splitting, can be implemented, in which molecules sharing the same scaffold are placed entirely in either the training set or the test set to more rigorously assess model generalisability. The training dataset is used to fit and train the model, while the test dataset is withheld from training and instead used to evaluate performance. During testing, the model receives only the input features and generates predictions, which are then compared with the true outputs. This procedure mimics real application, where the true output values are unknown.

As good practice, multiple iterations of train-test splits should be performed to ensure consistent performance. This is commonly achieved through *k*-fold cross-validation (CV), allowing several unique train-test splits, and several rounds of validation to be performed (**Figure 1.7**). The results from each round of validation can then be averaged to give an overall estimate of model performance.[93]

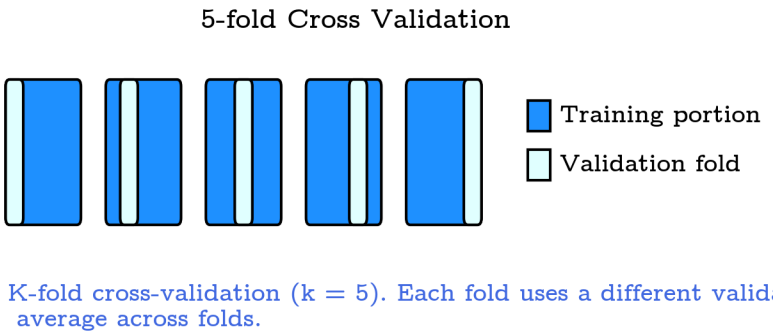


Figure 1.7: Illustration of 5-fold CV. The dataset splits into 5 equal folds. In each iteration, one fold (light blue) is used as the validation set while the remaining folds (dark blue) are used for training. This is repeated until every fold is used once as the validation set, the final performance is averaged across all folds.

Training the model involves adjusting the hyperparameters of the ML algorithm to minimise the error between predicted and true outputs, a process known as hyperparameter tuning. Hyperparameters are the parameters of an ML algorithm that are chosen before training and control how the model learns. The purpose of hyperparameter tuning is to tailor the algorithm to the specific problem. Different ML models have different sets of hyperparameters, and identifying the optimal combination is essential for achieving the best predictive performance.

The most common approaches to hyperparameter optimisation are grid search and random search. In both cases, a search space is defined where each dimension corresponds to a hyperparameter, and each point represents a possible configuration. In grid search, all points in the defined grid are systematically evaluated. In random search, points are sampled randomly from within predefined parameter bounds. The best-performing configuration is then selected, and the model is then fit on the entire training set through *k*-fold CV and evaluated on the test set.

k-fold CV helps diagnose overfitting as well as highlighting any selection bias. Overfitting, which is described in greater detail in **Section 1.3.6**, occurs when a model fits too specifically to the training data and fails to generalize to unseen data.[134] A telltale sign is strong internal CV performance on the training set but substantially poorer performance on the external test set. A large gap between train and test results

can also indicate selection bias or a distribution shift. Specifically, the CV folds could highlight an 'easy' subset of the training data by displaying overly optimistic results.

1.3.5 Model Evaluation

Once an algorithm has been fitted and optimised via hyperparameter tuning, the model's performance can then be evaluated. In regression models, one of the most widely used evaluation metrics is the mean absolute error (MAE), defined as the mean of the absolute differences between the predicted and true values (**Equation 1.1**). Another common metric is the root mean squared error (RMSE) (**Equation 1.2**). While RMSE penalises larger errors more heavily, MAE is often considered the most interpretable metric and can be used to compare model performance across both different models and datasets.[135]

$$\mathrm{MAE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1.1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1.2}$$

Where y_i and $\hat{y}_i$ are the experimental and predicted values for the i-th observation, respectively, and n is the total number of observations.

For classification models, the most reported metric is accuracy, defined as the ratio (or percentage) of correct predictions to the total number of predictions (**Equation 1.3**). Classification performance is frequently summarised in a confusion matrix, where the predicted classes are represented along one axis and the true classes along the other, providing a detailed breakdown of model accuracy across all classes (**Figure 1.8**).[136]

$$\mathrm{Accuracy} = \frac{(\text{Number of correct predictions})}{(\text{Total number of predictions})} \times 100 \tag{1.3}$$

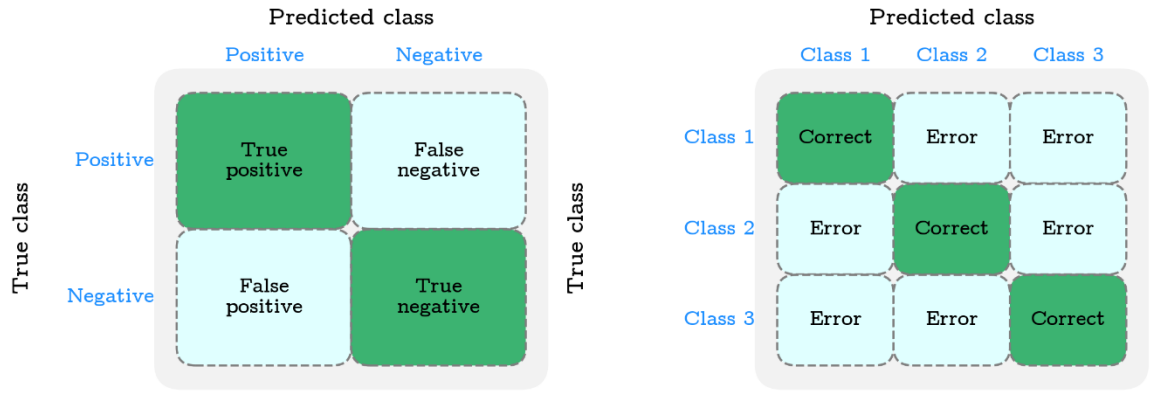


Figure 1.8: Confusion matrices for multiclass classification. (Left) 2x2 matrix with cells labelled for true positive, false negative, false positive and true negative. (Right) 3x3 matrix illustrating a multiclass setting. In both panels, rows are the true class and columns are the predicted class; diagonal cells (green) indicate correct predictions and off diagonal cells (light blue) indicate errors.

Beyond overall performance metrics, determining which features contribute most strongly to an ML model can yield valuable insight into the underlying factors that drive predictions. Permutation feature importance is a widely used method for assessing feature importance in a model-agnostic manner.[137, 138] In this approach, the trained model is first used to predict the outputs of the test set. Then, for each feature in turn, its values are randomly permuted (scrambled) while keeping all other features unchanged, and the model's performance is re-evaluated. The decrease in predictive performance indicates the relative importance of that feature. Repeating this process multiple times allows for the calculation of a mean importance score for each feature (**Equation 1.4**).

$$\text{Importance score} = \text{Accuracy (original)} - \text{Accuracy (permuted)} \tag{1.4}$$

Another useful procedure when analysing an ML model is the calculation of learning curves. In which, the model is refitted using increasingly larger subsamples of the training data and evaluating the performance on the test set. The resulting error can then be plotted against the size of the training set. This procedure provides information of the size of the dataset in terms of predictive performance, specifically if the same metrics could be obtained with a smaller number of data points. Conversely, if the curve has not plateaued and the metrics are still improving at the highest training set size, this can suggest a higher number of data points would improve performance.

1.3.6 Machine Learning Algorithms

While having the best dataset is arguably the most important part of an ML workflow,[139] the choice of model determines which patterns can be picked up from the data.

Supervised ML models must identify the essential structures and patterns within the training data while ignoring noise, minor fluctuations, and artefacts of data collection that lack chemical relevance for learning the objective function. When a model instead learns these irrelevant patterns or noise, overfitting occurs. Overfitting is particularly prevalent when the number of features per data point significantly exceeds the number of data points, as the model is more likely to memorise the dataset rather than learn the underlying patterns.[134]

The risk of overfitting also depends on the choice of model algorithm. Fundamentally, selecting the best ML algorithm for a given dataset is found by finding the best trade-off between bias and variance. In each ML algorithm different assumptions (biases) are made about the form of the target function. High-bias algorithms considerably simplify the relationship between inputs and outputs and are typically parametric. Because of their simplicity, they train quickly and require less data, but their limited flexibility can lead to underfitting. Low-bias algorithms, in contrast, are generally non-parametric, making fewer assumptions about the functional form and offering far greater flexibility. However, they involve more hyperparameters to optimise, require larger datasets, and demand longer training times.

Low-bias (flexible) algorithms are associated with high variance. Variance reflects how much a model's estimate of the target function would change if trained on different subsets of data. High-variance models may become overly sensitive to random fluctuations or noise, leading to overfitting and poor generalisability.[134] Conversely, high-bias models may underfit, as they lack the flexibility to capture the full complexity of relationships in the input features. This trade-off between bias and variance is illustrated in **Figure 1.9**.

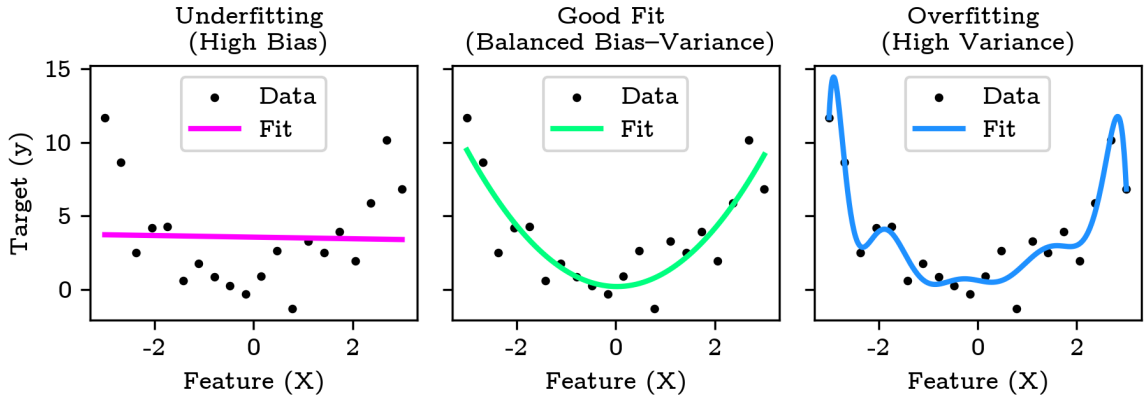


Figure 1.9: Illustration of the bias–variance trade-off in supervised machine learning. The left panel shows underfitting (high bias). The middle panel shows a good fit (balanced bias–variance). The right panel shows overfitting (high variance).

An overview of the models selected for use in this research is described in the following **Sections 1.3.7**, **1.3.8** and **1.3.9**. While this work focusses on these specific approaches, it is important to note that neural networks (NNs) have also achieved considerable success in chemistry,[140] including reaction outcome prediction.[141-143] Similarly, Bayesian optimisation (BO) has found recent applications within reaction optimisation workflows;[144] this thesis examines such an approach in Chapter 4.

1.3.7 Linear Regression Algorithms

Linear regression is one of the most widely used parametric algorithms, where a linear relationship is assumed between the input features and the output targets. The target function is shown in **Equation 1.5**:[145]

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (1.5)$$

Where:

$f(x)$ is the predicted output, x_1, x_2, ..., x_p are the input features, β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_p$ are the model coefficients (weights).

If only a single input feature is supplied, the model reduces to fitting a straight line. With multiple features, the model is termed multiple linear regression (MLR) and the relationship extends to a hyperplane in higher-dimensional feature space. The

parameters β are then estimated by minimising the least-squares error between the true outputs and the model predictions (**Equation 1.6**).

$$\min\beta \ \Sigma_i \ (y_i \ - \ f(x_i))^2 \qquad (1.6)$$

Where:

y_i is the observed (true) output value for sample *i*, $f(x_i)$ is the predicted output for sample *i* based on the linear model, β is the set of model coefficients being estimated and Σ_i is the summation that runs over all training examples.

Large coefficients can make the model unstable and prone to overfitting. To address this, regularisation terms are introduced that penalise large parameter values. Different variants of linear regression models employ different forms of regularisation. The least absolute shrinkage and selection operator (LASSO) regression (L1 regularisation) penalises the sum of the absolute values of the coefficients.[146] This can shrink less important coefficients exactly to zero, effectively performing feature selection. Ridge regression (L2 regularisation) adds a penalty proportional to the sum of squared coefficients, shrinking all coefficients towards zero but never eliminating them entirely.[147] An illustration of both LASSO and ridge regression is shown in **Figure 1.10**.

In both cases, the strength of the penalty is controlled by the hyperparameter λ. When λ=0, the model reduces to ordinary least squares regression, while larger values of λ increase the influence of the penalty.

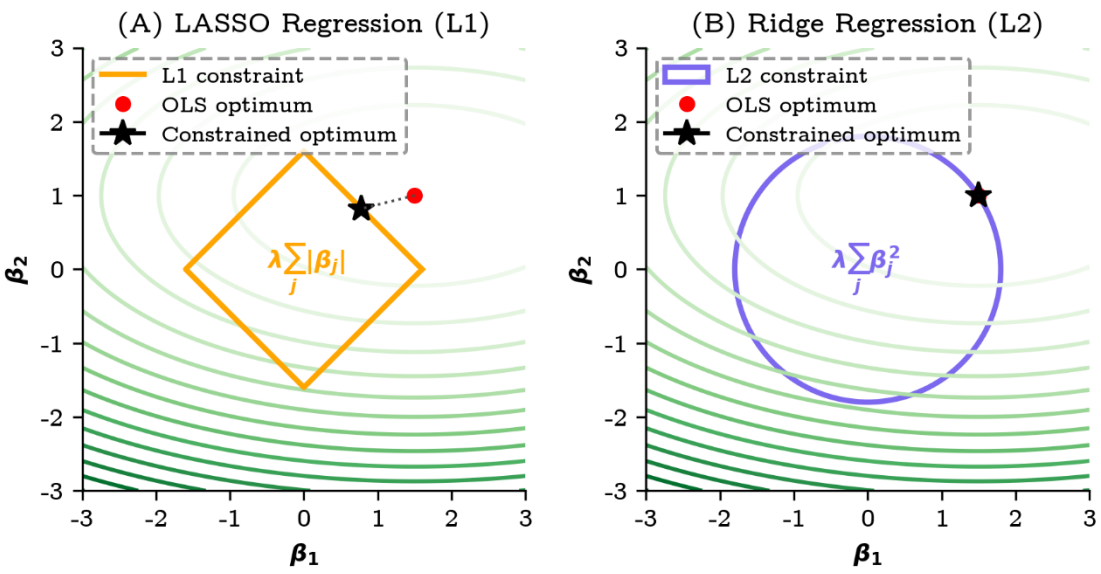


Figure 1.10: Illustration of A) LASSO and B) Ridge regression. The diamond region in orange and the circle in purple show the constrained region for LASSO and Ridge regression respectively. The constrained optimum is a black star and the equation for the penalty term used in each regression model is shown in the middle of each graph.

1.3.8 Decision Tree Algorithms

In contrast to linear regression algorithms, which are typically only applied to regression tasks, decision tree algorithms can be applied to both classification and regression problems. A decision tree consists of a hierarchy of compiled nodes that split the data depending on the values of input features. Each internal node branches into another two nodes that split the data further; this is carried out until the end of the tree is reached, where the last nodes are termed the leaf nodes.

In classification tasks, trees are trained so that each leaf node corresponds to a particular class. In regression tasks, nodes are created to minimise the RMSE between the model's predictions and the true output values.

Several widely used ML algorithms are based on ensembles of decision trees, including random forests, extremely randomised trees (extra-trees) and gradient boosting.

The random forest algorithm is the most common and has been used successfully for both regression and classification predications of problems.[148-151] In this algorithm multiple independent, unpruned decision trees are trained using the bootstrap aggregation (bagging) technique.[138] In bagging, each tree is trained on a

different bootstrap sample of the training data, generated by sampling with replacement.[152] At each node, an optimised specific subset of input features is selected for splitting. This design encourages diversity among trees so that they make different predictions with different errors. The overall prediction of the ensemble is obtained by averaging (regression) or majority voting (classification) across all trees. Pruning, by contrast, is a separate complexity reduction technique in which branches of a tree are removed if they are not critical for prediction.

The extra-trees algorithm is closely related to random forests.[153] However, unlike random forests, each tree is trained on the full dataset rather than a bootstrap sample. Diversity is introduced by splitting features randomly at each node, rather than optimising splits. This approach is computationally less expensive than random forests while still maintaining diversity in the ensemble. An illustration of random forest and extra-trees is shown in **Figure 1.11**.

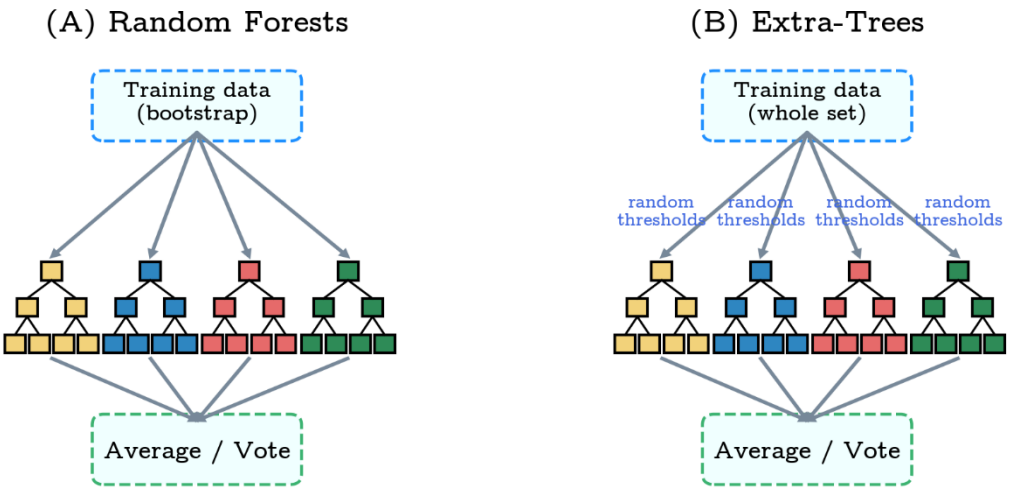


Figure 1.11: Illustration of Random Forests and Extra-Trees algorithm. A) Random Forests: multiple decision trees are trained on different bootstrap samples of the training data, with random subsets of features at each split. B) Extra-Trees: all decision trees are trained on the entire dataset, but split thresholds are chosen randomly. In both cases the output is obtained from the average predictions of all individual trees.

Gradient boosting differs from random forests and extra-trees in that trees are built sequentially rather than independently (**Figure 1.12**).[154, 155] Boosting refers to methods that transform weak learners (such as decision trees) into a strong predictive model through iterative training. In gradient boosting, each new tree is trained to reduce the

residual error from the ensemble of previously built trees.[156] Although this process is typically slower to train, it often results in superior predictive performance.

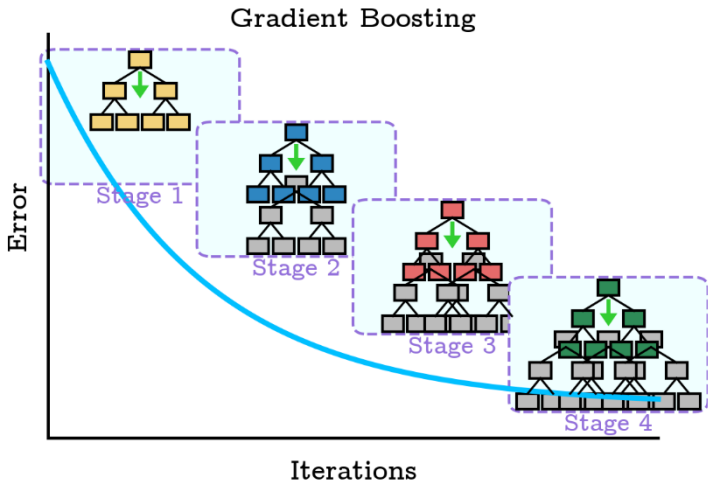


Figure 1.12: Schematic illustration of Gradient Boosting. At each stage of training each new tree (coloured) is fit to the residual errors of the current ensemble (grey), gradually reducing the overall training error (blue curve).

1.3.9 Support Vector Machines

Support Vector Machines (SVMs) are a class of supervised learning algorithms that can be applied to both classification and regression tasks, support vector classification (SVC) and support vector regression (SVR).[157] The central concept underlying SVMs is the kernel trick, which enables data to be implicitly mapped into a higher-dimensional feature space.[158] This approach is also central to other kernelized methods such as Gaussian Process Regression (GPR). The approach allows linear models in the new higher-dimensional space to represent highly complex, nonlinear relationships in the original input space. Crucially, the kernel trick avoids the explicit computation of coordinates in the higher-dimensional space, making such models computationally efficient. Common kernel functions include the linear, polynomial, sigmoid, Laplacian, and Radial Basis Function (RBF) kernels, all of which may be applied to either classification or regression variants of SVMs.

In classification tasks, SVMs construct a decision boundary (a hyperplane in the transformed feature space) that separates data points belonging to different classes. The algorithm seeks the hyperplane that maximizes the margin, defined as the distance

between the boundary and the closest data points of each class. These closest points are the support vectors, which determine the orientation and position of the hyperplane. The kernel function enables SVC to capture nonlinear class boundaries by implicitly operating in higher dimensions. Predictions are then made based on which side of the decision boundary a new point lies.

In regression tasks, the same principles are adopted but the objective differs. Instead of separating classes, SVR aims to find a function that approximates the continuous output values within a specified tolerance. This tolerance is defined by the ε-insensitive loss function, which introduces a tube (the ε-tube) around the regression function. Errors that fall within this tube are ignored, while deviations outside are penalized. Only the data points lying outside or on the boundary of the ε-tube influence the model parameters, these are the support vectors in the regression setting. By combining the ε-insensitive loss with kernel functions, SVR can model both linear and highly nonlinear regression problems. An illustrative example of SVC and SVR is presented in **Figure 1.13**.

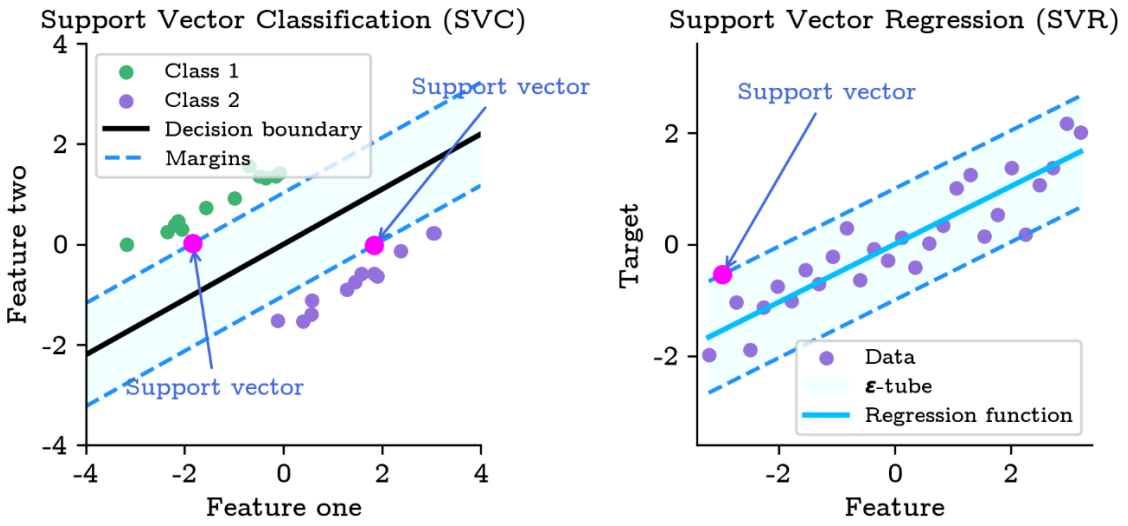


Figure 1.13: Schematic representation of Support Vector Machines for classification (left) and regression (right), illustrating the separating hyperplane or regression function together with their respective margins or ε-tube, class data points, and support vectors.

1.4 MACHINE LEARNING FOR REACTION OUTCOME PREDICTION

ML has been applied in many areas of chemistry, including reaction optimisation through the prediction of reaction outcomes,[26, 159] such as reaction yields,[68, 70, 160-162] and, most notably for the purpose of this work, the prediction of enantioselectivities in asymmetric transformations.[105, 163-168] Predicting experimental outcomes, however, remains inherently challenging because chemical reactions typically involve multiple molecular components and parameters. Consequently, the input features used in ML models must capture this complexity.[169] Broadly, ML approaches to reaction outcome prediction can be distinguished by whether or not they rely on mechanistic knowledge to generate their input features.

Mechanism-based approaches typically involve QM/MM calculations of key transition states (TSs) in the catalytic cycle.[170] From these, specific physical-organic features (as discussed in **Section 1.3.3**) can be extracted. These models are computationally intensive and tend to be specific to the catalytic system being studied. Only a limited number of studies have applied this strategy to AH reactions, and those reported to date focus specifically on Rh-based systems.[166, 171, 172]

An alternative, mechanism-free strategy is the use of quantitative structure–property relationships (QSPR),[173, 174] which aim to find correlations between chemical structure and performance. ECFPs are a common example of such descriptors.

A hybrid approach, integrating computationally derived descriptors with QSPR models, has also been used. In some cases, specific features can be extracted from ground-state structures of reactants, products, and catalysts, avoiding the high computational cost of obtaining TS energies. This is particularly advantageous for AH of TSC=C, where generating a sufficiently large and diverse dataset of TS energies for such a complex reaction pathway is both computationally prohibitive and impractical.

Beyond the type of descriptors used another distinction of ML approaches for reaction outcome prediction is the source of training data. Datasets can be obtained from HTE, computational studies, conventional batch experiments, or manual curation of literature reports (as discussed in **Section 1.3.2**).[101]

Recent efforts have successfully applied ML to predict reaction outcomes of yield and enantioselectivity. For example, the Doyle group developed ML models to predict experimental yields in the Pd-catalysed Buchwald-Hartwig amination of aryl halides with 4-methylaniline.[68] In this work, an experimental dataset consisting of 4,014 reactions was created through HTE, and physical-organic features were generated from DFT-derived ground state structures of the reaction components. Using a random forest regressor (RFR), the model achieved a coefficient of determination (R^2) value of 0.92.

In contrast, the Denmark group reported ML models for predicting enantioselectivity in chiral phosphoric acid–catalysed asymmetric *N*,*S*-acetal formation.[105] In this case, features were derived from DFT-calculated TS energies corresponding to the competing enantiomeric pathways, generating a combinatorial dataset of 1,075 reactions. SVR models were trained to predict the difference in activation free energy between the two competing TSs ($\Delta\Delta G^{\ddagger}$) values rather than experimental *ee*, with only 15 catalysts tested experimentally for calibration and validation. The trained models achieved an MAE of 0.152 kcal mol^{-1} and were subsequently applied to an in-silico library of 1,000 candidate catalysts. From this screen, three catalysts predicted to give the highest *ee* values were synthesized, all three outperformed the original experimental set.

Building on this, the Glorius group introduced a multiple fingerprint feature (MFF) model, in which a deep neural network (DNN) was trained directly on structural fingerprints of the Denmark dataset (1,075 reactions).[175] Using these mechanism-agnostic descriptors, they reported a slightly higher predictive result, achieving an MAE of 0.144 kcal mol^{-1}, thereby matching or surpassing the TS-based model while avoiding costly DFT calculations.

The competitive results seen from the Glorius group demonstrate that reactivity can, in some cases, be predicted directly from chemical structures using fingerprints alone, thereby avoiding the need for computationally-expensive DFT TS calculations. A similar point was underscored by Doyle's work on yield prediction, where strong

predictive performance was achieved using features derived solely from optimised ground-state structures.

1.4.1 Machine Learning for *ee* Prediction in AH Reactions of Olefins

To date, only two studies have reported the application of ML approaches to predict the experimental *ee* of AH reactions of olefins.

In 2019 Singh *et al*,. reported a proof-of-concept study applying ML to predict *ee* values for AH reactions of alkenes and imines.[148] Their protocol focussed exclusively on five families of asymmetric binaphthyl-derived catalysts. A dataset comprising 368 known AH reactions were compiled and subdivided into five groups according to the catalyst family employed. Physical-organic descriptors were generated for the ML models from DFT-optimised ground states of the chemical structures in the dataset (including HOMO and LUMO energies, bond lengths, tPSA, and related features).

ML models were trained independently on each subset using RFR, convolutional neural networks (CNN),[176] gradient boosting regression (GBR), and linear regression. Among these, RFR exhibited the best performance, with RMSEs of 6.3, 6.5, 9.2, 8.2, and 7.1% *ee* for the respective subsets. The subsets were subsequently merged into a larger combined dataset containing all reactions, yielding an overall RMSE of 8.4 $\pm$ 1.8% *ee*. SMOTE was employed to address the class imbalance that existed in the dataset,[103] however this adjustment did not significantly improve model performance.

Notably, the RMSE obtained from the combined dataset overlapped with those derived from the individual subsets, indicating the structural diversity that the RFR algorithm could accommodate. Nevertheless, it should be emphasised that binaphthyl-derived catalysts represent only a narrow fraction of the asymmetric catalysts employed in AH reactions. Moreover, the dataset did not include any examples of tetra-substituted alkenes, limiting its broader generality.

ln 2021 Hong *et al.*[25] developed an ML procedure to find the optimum asymmetric catalyst for the AH of a target olefin. The experimental data from 12,619 AH reactions on 2754 di-, tri- and tetra-substituted olefins collected from the literature, and extended tight-binding (xTB) optimised geometries were generated for the

reactants, products, and catalysts. The resulting dataset provides a valuable resource for ML applications in this domain.

The procedure outlined followed a hierarchical learning strategy.[177] For each olefin to be targeted, different ML models would be built from different subsets of the original dataset. Beginning with the target substrate, a subset was formed of reactions of olefins that matched one substituent of the target olefin. This dataset was used to form a base model to make base predictions of the *ee*. A second dataset was subsequently formed where two substituents matched that on the target olefin, this dataset was used to learn the difference between the reported *ee* and the predicted values from the base model for the substrates in this second dataset, a process known as Δ learning.[177] For tri-substituted olefins this step could be repeated with a third dataset being formed of the olefins that match three of the substituents on the target olefin. The final prediction is the sum of the predicted values from the base model and the Δ model(s). This publication utilised the most common olefin in the original dataset ((Z)-2-acetamido-3-phenyl-acrylate) for this procedure.

The features input into the base and Δ models included both physical-organic (3D features) and molecular fingerprints (2D features). The models were built on 2D and 3D descriptors independently and combined. The best performance was found with the combination of 2D and 3D features with the Extra Trees algorithm for the olefin tested.

The hierarchical workflow performed well on the most common olefin; however, it has not been evaluated on less common or entirely novel substrates for which little or no experimental data are available. Consequently, its performance in these cases remains unclear.

1.5 PROJECT AIMS

The overarching aim of this research was to develop ML models capable of predicting the *ee* of AH reactions of TSC=C, thereby reducing the reliance on costly trial-and-error experimentation traditionally employed in optimisation campaigns.

The first objective (Chapter 2) was to assemble a comprehensive dataset of all available TSC=C reactions reported in the literature, to enable development of ML models trained on reaction features to predict the *ee* of AH reactions of TSC=C, and to evaluate their predictive ability by comparing predicted and experimental *ee* values.

The third objective (Chapter 3) was to assess the performance of novel AH reactions of TSC=C, by conducting reactions experimentally and comparing the predicted *ee* values to the experimentally determined results. This aimed to evaluate the potential of the models for integration into real experimental workflows, which would enable synthetic chemists to more efficiently identify highly enantioselective reactions and reduce the number of experiments required in a typical reaction optimisation study.

Finally, the fourth objective (Chapter 4) was to explore the application of Bayesian optimisation (BO) to guide the AH of a challenging TSC=C substrate, using the workflow to identify in real time the most enantioselective and high-yielding reactions from a predefined set of candidates.

2 Using Machine Learning to Predict Enantiomeric Excess of a Literature Dataset of Asymmetric Hydrogenations of Tetra-substituted Olefins

2.1 INTRODUCTION

2.1.1 Chapter Aims

The aim of this project was to develop ML models capable of predicting the *ee* of AH reactions of TSC=C reported in the literature. This was accomplished by: (i) assembling a comprehensive dataset of all relevant literature examples of AH of TSC=C; (ii) carrying out conformational searches and geometry optimisations for the chemical structures in the dataset, specifically the reactants, products and metal catalysts; (iii) generating input features for each reaction that were both suitable for ML model input and informative for predicting *ee*; (iv) constructing and training a series of ML models to predict the *ee* of the compiled dataset; and (v) evaluating the models in terms of predictive performance and the most informative input features.

2.1.2 Computational Theory

As outlined in chapter 1, many effective input features for ML are extracted from computationally optimised molecular structures. In this project, such features were generated to evaluate their predictive power for the *ee* of the studied reaction class. Since the QC, electronic, and steric properties of a structure vary across conformers, it is essential to identify the conformer corresponding to the global minimum, the lowest-energy ground state on the potential energy surface (PES), to obtain reliable parameters.[178]

The PES describes the potential energy of a chemical system in terms of a function of its coordinates. Reactants and products are found as minima on the PES due to their stable chemical nature, while saddle points are high in energy and correspond to TSs. The PES typically contains many local minima, the lowest of which is called the global minimum and corresponds to the true ground state chemical structure. This global minimum can be found through conformational searching and subsequent geometry optimisation techniques, utilising both QC and MM methods.

Conformational searching explores the PES by generating multiple conformers of a chemical system to identify the lowest-energy structure. Starting from an initial geometry, new conformers are produced iteratively, and searches are restarted

whenever a lower-energy conformer is discovered. Conformers are distinguished by their relative energies, rotational constants, and root mean squared deviations (RMSD). Geometry optimisation methods optimise the conformer to the lowest energy chemical arrangement that can be found, the local minimum. If conformational searching has successfully found the lowest energy conformer, this local minimum will in fact be the global minimum

Geometry optimisation can be performed using a range of QC methods, including DFT and semi-empirical quantum mechanical (SQM) approaches. DFT is a widely used method in which the ground-state electronic energy is determined entirely by the electron density,[179, 180] offering good accuracy for electronic structures and energies. However, DFT can become computationally demanding for larger systems and for those containing transition-metals. In contrast, SQM methods introduce approximations and incorporate parameters derived from experimental data or high-level theoretical calculations. This makes them significantly faster and suitable for larger systems, though generally at the cost of lower accuracy.[181] A widely adopted subset of SQM methods is based on DFT, known as density functional tight binding (DFTB) methods. Examples include the geometry, frequency, noncovalent (GFN) method, GFNn-xTB (n=0,1,2).[182, 183]

Traditionally, conformational searches are performed using MM methods. Most MM approaches rely on Monte Carlo random sampling to generate conformers.[184, 185] The energies are calculated using predefined force fields, which describe atomic interactions based on parameters derived from experiment. While these force fields are transferable between many systems, they are not universally applicable. With the exception of the Universal Force Field (UFF),[186] force fields have not yet been parametrized for the whole of the periodic table, which limits the applicability of MM methods to systems containing elements such as 3d–5d transition metals.[187]

To address this limitation, QC based alternatives have been developed. One such method is the Conformer–Rotamer Ensemble Sampling Tool (CREST), which combines RMSD-based metadynamics simulations with the SQM GFNn-xTB methods.[188] Conformers are generated using the iMTD-GC algorithm,[189] an iterative

approach that integrates metadynamics (MTD) sampling to explore the PES with genetic cross-over (GC) techniques to promote structural diversity.[190] Each conformer is subsequently optimised at the GFNn-xTB (n = 0, 1, or 2) level of theory.

Alternatively, the iMTD-GC algorithm can be combined with geometry optimisation at the GFN-FF level.[191] GFN-FF is a generic force field parameterized to reproduce geometries obtained using the density functional B97-3c.[192] Importantly, both GFNn-xTB and GFN-FF are broadly parameterized for 3d–5d transition metals,[187] offering significant advantages over UFF and demonstrating strong performance in the literature.[193-197]

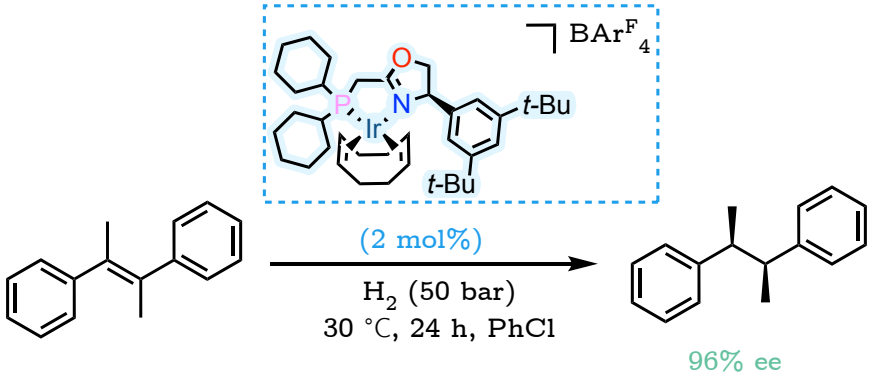
2.2 METHODS

2.2.1 Dataset Generation (i)

The dataset for this project was compiled from published chemical studies reporting AH reactions of TSC=C. In 2021, Hong *et al.* published a database containing 12,619 reactions involving di-, tri-, and tetra-substituted olefins, extracted from 355 publications.[25] Of these, 41 papers reported reactions specifically involving TSC=C, providing a valuable starting point for dataset construction. An additional 27 relevant papers were subsequently identified, yielding a final dataset of 1,271 reactions.[4, 8, 29, 33, 36, 38, 40-45, 198-249]

For each reaction, the following information was recorded: reactant, product, metal employed, catalyst, temperature, pressure, additive, solvent, catalyst loading S/C), scaffold-type, *ee* and the Digital Object Identifier (DOI) for the paper for which the reaction came from. An example of a reaction in the dataset is shown **Table 2.1**. Chemical structures were recorded as their SMILES representations, with ancillary ligands removed from metal catalysts. As a given structure can be represented by multiple SMILES strings, standardization was performed using the rational discovery kit (RDKit). Additionally, as SMILES lack a built-in method to encode metal–ligand coordination, a custom script was implemented to represent dative bonds from coordinated ligands to the metal centre.

Table 2.1: Example row from the dataset of AH reactions of TSC=C extracted from the literature. The highlighted region of the metal catalyst corresponds to the portion recorded in SMILES format, with ancillary ligands removed. The complete dataset contains 1,271 reactions



Subst.	Prod.	Cat.	Lig.	Metal	Scaff-old	Ad.	Solv.	Pressure H_2 (bar)	Temp. (°C)	Cat. load (S/C)	ee	DOI
C/C(=C(/C)c1ccccc1)c1ccccc1	C[CH](c1ccccc1)[C@H](C)c1ccccc1	CC(C)(C)c1cc([C@@H]2COC3=N2->[Ir+]<-P(..	CC(C)(C)c1cc([C@@H]2COC(P...	Ir	P, N type		Clc1ccccc1	50	30	50	0.96	10.1002/ange.201912640

2.2.2 Computational Methods (ii)

From the SMILES representations, 3D coordinates were generated for each structure in the dataset using RDKit.[124] Conformational searches and geometry optimisations of the ground state structures in the dataset were then performed with CREST at the GFN2-xTB level of theory.[188] During longer simulations of ferrocene-containing structures, incorrect changes in connectivity were observed in 13 of the 50 ferrocene-containing metal catalysts. For these cases, conformational searches were instead carried out with GFN-FF,[191] followed by geometry optimisation with GFN2-xTB.

2.2.3 Feature Generation (iii)

To make the data interpretable for ML algorithms, reaction conditions, including solvent, additives, metal, and scaffold type, were converted into numerical format. The chemical structures of the reactant, product, and catalyst were represented by ECFPs generated from their corresponding 3D coordinates using RDKit, with a radius of three and a fixed fingerprint length of 2,048 bits.[116]

In addition, a range of physical–organic descriptors were extracted for each structure in the dataset from the optimised ground state structures using various Python toolkits, **Table 2.2**. Features were computed at both the molecular level and, where applicable, for selected common atoms. For this reaction system, the common atoms include: the reaction centre of the reactant and product (the two carbons of the double bond, C1 and C2), the metal centre in the catalyst (M), and the two atoms of the ligand coordinated to the metal centre (L1 and L2, **Figure 2.1**). Notably, all metal catalysts in this dataset consisted of a single metal centre with bidentate ligands.

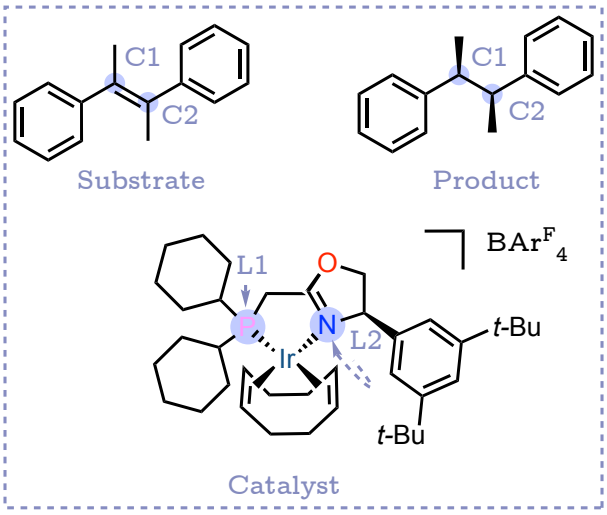


Figure 2.1: Example of the common atoms in chemical structures from the dataset, highlighting those for which features were extracted at the atomic level.

Table 2.2: Features extracted for ML including a) reaction conditions extracted from literature publications; b) physical-organic features extracted for the reactant, product, and metal catalyst at the molecular level; c) physical-organic features extracted for the reactant, product and metal catalyst for the common atoms: M, C_1, C_2, L_1 and L_2; d) ECFPs extracted for each reactant product and metal catalyst at radius 3 at a fixed length of 2048 bits.

Feature	Source	Description
Energy[b]	Goodvibes[250]	Electronic Energy
Enthalpy[b]	Goodvibes	Enthalpy
QH Entropy[b]	Goodvibes	Quasiharmonic energy
ZPE[b]	Goodvibes	Zero-point energy
HB Acceptors[b]	Pybel[251]	Number of hydrogen bond acceptors
HB donor[b]	Pybel	Number of hydrogen bond donors
TPSA[b]	Pybel	Topological polar surface area
Chemical Pot.[b]	Cclib[252]	Global chemical potential
HOMO[b]	Cclib	Highest occupied molecular orbital (HOMO) energy
LUMO[b]	Cclib	Lowest unoccupied molecular orbital (HOMO) energy
Vib Freq.[b]	Cclib	Lowest vibrational frequency
Vib IR[b]	Cclib	Lowest infrared intensity
Mulliken[b,c]	Cclib	Muliken atom charge
SASA[b,c]	DBSTEP[253]	Solvent accessible surface area
PBV[b,c]	DBSTEP	Percent buried volume (3.5 Å radius)
Sterimol B1,B5, L[b,c]	DBSTEP	Sterimol Bmin, Bmax, and L parameter[130]
PEOE[b,c]	RDKit[124]	Partial equalisation of orbital electronegativities
LogP[b,c]	RDKit	Wildman-Crippen partition coefficient
MR[b,c]	RDKit	Wildman-Crippen molar refractivity
Estate[b,c]	RDKit	Electrotopological state index
MW[b]	RDKit	Molecular weight
Chi[b]	RDKit	Molecular connectivity Chi indexes[254]
Ipc[b]	RDKit	Information polynomial coefficient[255, 256]
BalabanJ[b]	RDKit	Balaban topological index[257]
BertzCT[b]	RDKit	Bertz index of molecular complexity[258]
Kappa[b]	RDKit	Kappa shape indexes[254]
NHOHcount[b]	RDKit	Number of NHs or OHs
NOcount[b]	RDKit	Number of nitrogens and oxygens
ECFP[d]	RDKit	Extended connectivity molecular fingerprint
Metal[a]	Literature	Metal present in the metal catalyst
Additives[a]	Literature	Additives present in the reaction
Scaffold type[a]	Literature	Type of ligand scaffold in the metal centre
Temperature[a]	Literature	Temperature of the reaction
Pressure[a]	Literature	Pressure of the reaction
Solvent[a]	Literature	Solvent the reaction was carried out in
S/C[a]	Literature	Substrate over catalyst concentration

2.2.4 Machine Learning Methods (iv)

The Python ML library scikit-learn (sklearn) was used for dataset splitting, model construction, training, and evaluation. An 80:20 train–test split was applied to the dataset prior to model training, resulting in a training set of 1,017 reactions and a test set of 254 reactions. Both regression and classification supervised learning models were employed, specifically: Lasso linear regressor (LLR), ridge regressor (RR), random forest regressor (RFR), gradient boosting regressor (GBR), extra-trees regressor (ETR), random forest classifier (RFC), gradient boosting classifier (GBC), and extra-trees classifier (ETC).

Each model was trained and evaluated using three subsets of input features, summarised in **Table 2.3**. For clarity, these are referred to as 2D descriptors (subset 1), 3D descriptors (subset 2), and their combination, 2D + 3D descriptors (subset 3). The 2D descriptors consist of ECFPs derived from chemical structures, while the 3D descriptors comprise the physical–organic features (**Table 2.2**) extracted from optimised structures in the dataset. All feature sets additionally include the seven features describing the reaction conditions which include, scaffold type, temperature, pressure, solvent, catalyst loading, additives and metal centre.

Table 2.3: The three different subsets of input features employed in this project. The reaction conditions are a set of 7 features including scaffold type, temperature, pressure, solvent, catalyst loading, additives and metal centre. The physical-organic features are composed of 250 individual features described in Table 2.2.

Input Type	Description	No. Input Features
(2D)	Reaction conditions + ECFPs of chemical structures	6,151
(3D)	Reaction conditions + physical-organic features	257
(2D + 3D)	Reaction conditions + ECFPs + physical-organic features	6,401

Prior to model training, the input features were standardized using sklearn's StandardScaler. Hyperparameter optimisation was then performed on the training set with scikit-learn's GridSearchCV, searching for the combination of hyperparameters that yielded the highest MAE or accuracy score under 5-fold CV. The hyperparameter search space for each model is detailed in **Table 2.4**. Each algorithm was subsequently retrained on the full training set using its optimal

hyperparameters (Appendix), before being applied to predict the outputs of the test set.

Table 2.4: The sklearn hyperparameters tuned for each model algorithm. The ML algorithms based of decision trees including random forest, extra-trees and gradient boosting. The support vector regression model was fit with the RBF kernel. Model abbreviations are defined as follows: LLR—Lasso Linear Regressor, RR—Ridge Regressor, RFR—Random Forest Regressor, RFC—Random Forest Classifier, GBR—Gradient Boosting Regressor, GBC—Gradient Boosting Classifer, ETR—Extra-Trees Regressor, ETC—Extra-Trees Classifier, SVR—Support Vector Regressor.

Model	Hyperparameters	Values
LLR	α	1e-06, 1e-05, 0.0001, 0.001, 0.001, 0.1, 1, 10, 100
RR	α	1e-06, 1e-05, 0.0001, 0.001, 0.001, 0.1, 1, 10, 100
RFR *RFC*	Max_features	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
	Max_samples	0.001, 0.01, 0.1, 1
	Min_samples_leaf	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
	Min_samples_split	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
	Max_depth:	None, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20
	n_estimators	200, 300, 400, 500, 600, 700, 800, 1000, 1200, 2000
GBR *GBC*	Learning_rate	0.001, 0.01, 0.05, 0.1, 1
	Max_depth	None, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20
	Max_features	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
	Min_samples_leaf	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
	Min_samples_split	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
	n_estimators	200, 300, 400, 500, 600, 700, 800, 1000, 1200, 2000
ETR *ETC*	Max_features	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
	Max_depth	None, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20
	Min_samples_leaf	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
	Min_samples_split	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
	n_estimators	200, 300, 400, 500, 600, 700, 800, 1000, 1200, 2000
SVR	Gamma	scale, auto
	C	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20
	Epsilon	1e-10, 1e-09, 1e-08, 1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5

2.2.5 Machine Learning Evaluation (v)

To assess the performance of each ML regression model, the *ee* values for the test set were compared with the corresponding experimental values, and the MAE between them was reported (**Equation 1.1**).

For the classification models, the target *ee* values were categorised into predefined classes based on their magnitude. The exact classes used will be discussed in **Section 2.3.3**. The models were then trained to predict the class membership of

each instance rather than its exact *ee* value. Model performance was then evaluated using classification accuracy, defined as the percentage of predictions falling into the correct class (**Equation 1.3**).

Model performance was assessed using 5-fold CV. For each fold, MAE and accuracy were computed on both the training and test sets. The mean ± SD across folds of the train and test sets were reported.

To validate the adequacy of the training set, identify potential overfitting, and assess whether equivalent performance could be achieved with less data, learning curves were generated for each fitted model. This was achieved by refitting the model on progressively larger fractions of the training data and plotting the resulting training and test MAE and accuracy as a function of training-set size.

Permutation feature importances were also calculated for the models, by permuting each feature in the training set and measuring the change in MAE or accuracy on the test set, as discussed in **Section 1.3.5**.

2.3 RESULTS AND DISCUSSION

2.3.1 Dataset Analysis

The dataset used for this ML project comprised 1,271 AH reactions of TSC=C, extracted from 68 publications. Across these reactions, 317 distinct metal catalysts and 468 unique substrates were identified. The most prevalent substrates ranked by publication count are shown in **Figure 2.2**, while those ranked by total reaction count are shown in **Figure 2.3**. When ranked by publication count, the figure reflects structures that are widely studied across the literature, even if they appear only a few times in each paper. These substrates likely represent broadly applicable or benchmark substrates for AH reactions, tested by multiple research groups under varied reaction conditions. In contrast, when ranked by reaction count, the list shifts and reflects frequently repeated structures that are concentrated in just one or two publications, indicating they are likely substrates within optimisation campaigns rather than benchmark substrates.

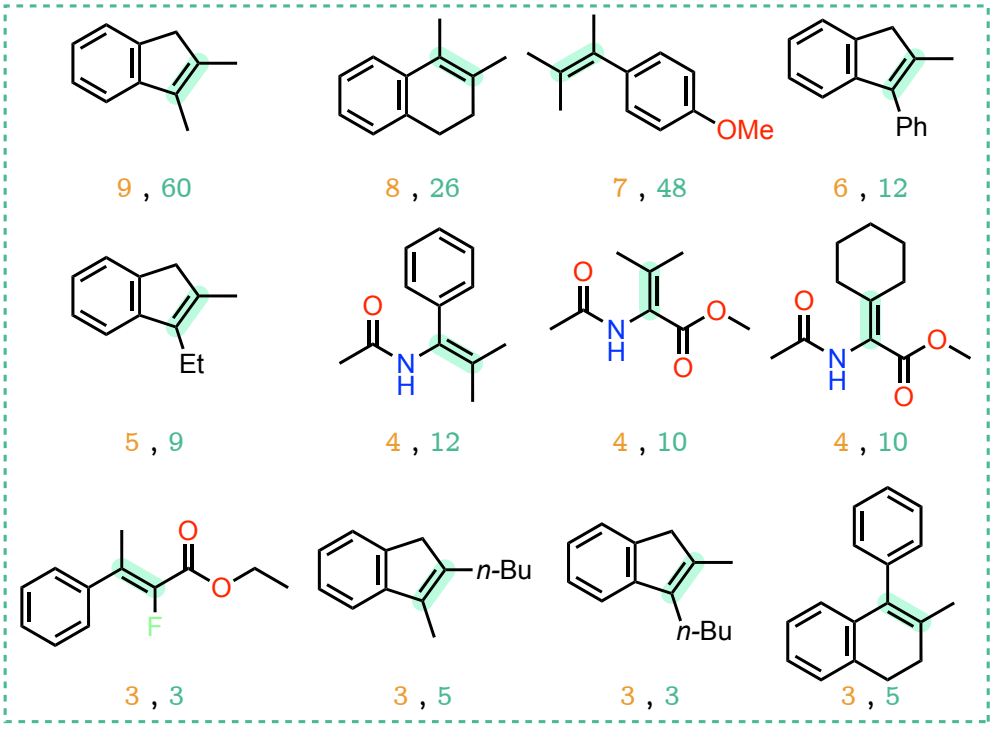


Figure 2.2: Chemical structures of the 12 most frequently reported substrates in the dataset, ranked by the number of unique publications (DOIs) in which they appear. The number of publications for each substrate is shown below in orange, and the corresponding number of reactions is shown in green.

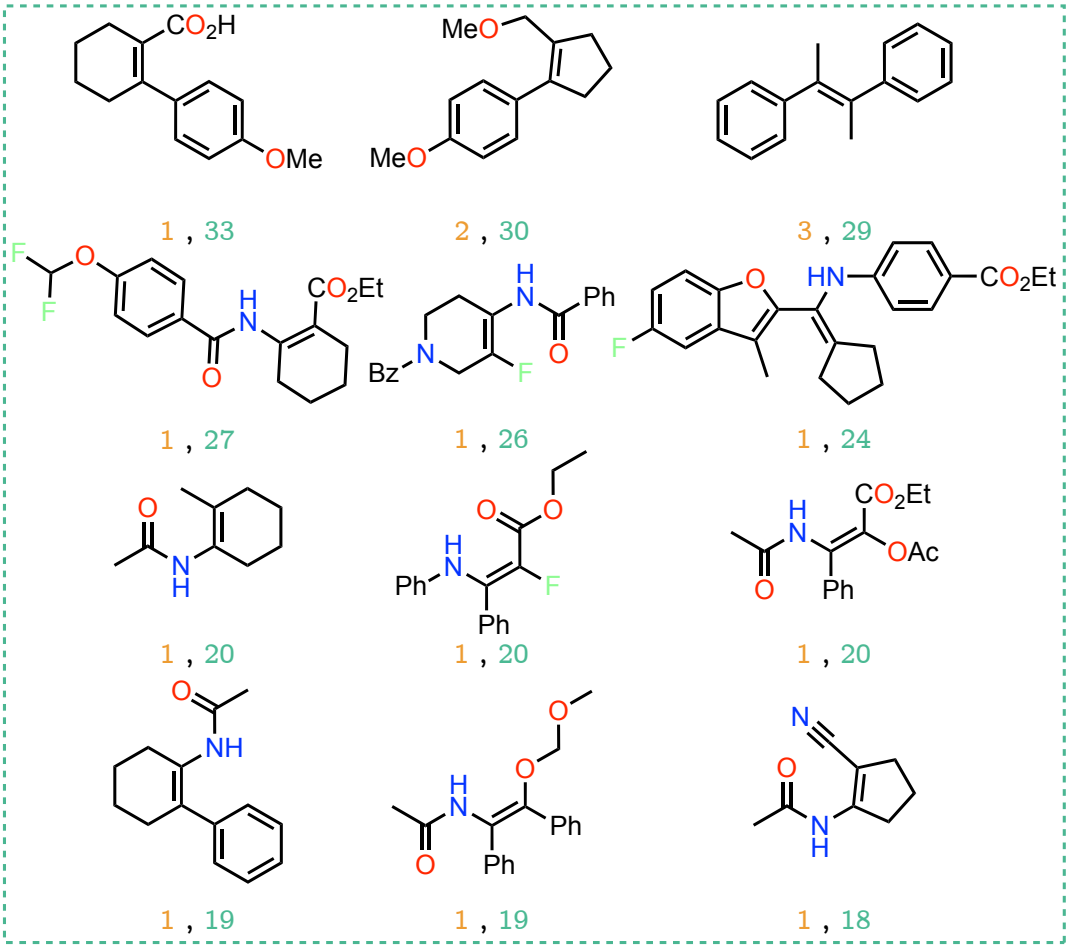


Figure 2.3: Chemical structures of the 12 most frequently reported substrates in the dataset, ranked by the number of reactions in which they appear. The number of publications for each substrate is shown below in orange, and the corresponding number of reactions is shown in green.

A similar analysis was performed for chiral ligands, ranking them both by the number of unique publications and by total reaction count (**Figures 2.4** and **2.5**). Ranking by publication count highlights ligands that have been applied across a wide range of AH studies and substrates and are therefore well-established ligand systems for this reaction. Ranking by reaction count identifies ligands that are associated with one or two publications, reflecting their use in targeted optimisation campaigns.

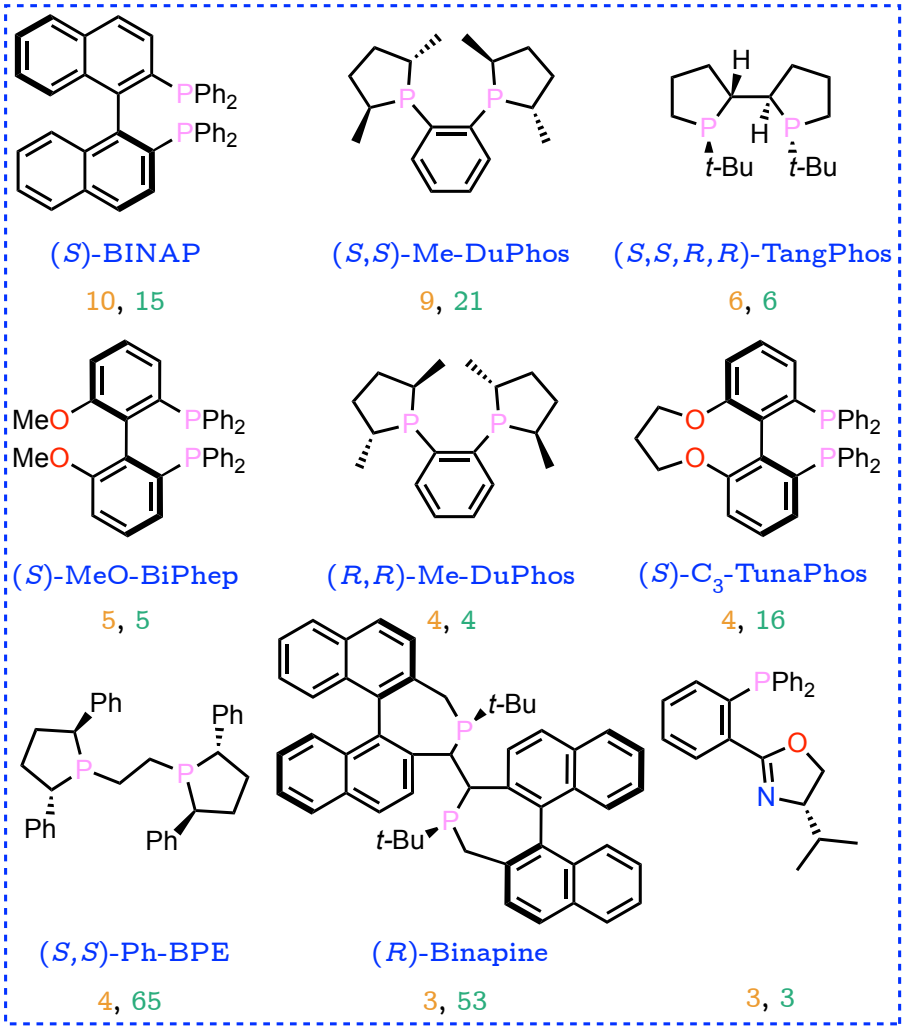


Figure 2.4: Chemical structures of the nine most frequently reported ligands in the dataset, ranked by the number of unique publications (DOIs) in which they appear. Publication counts are shown in orange, and reaction counts in green.

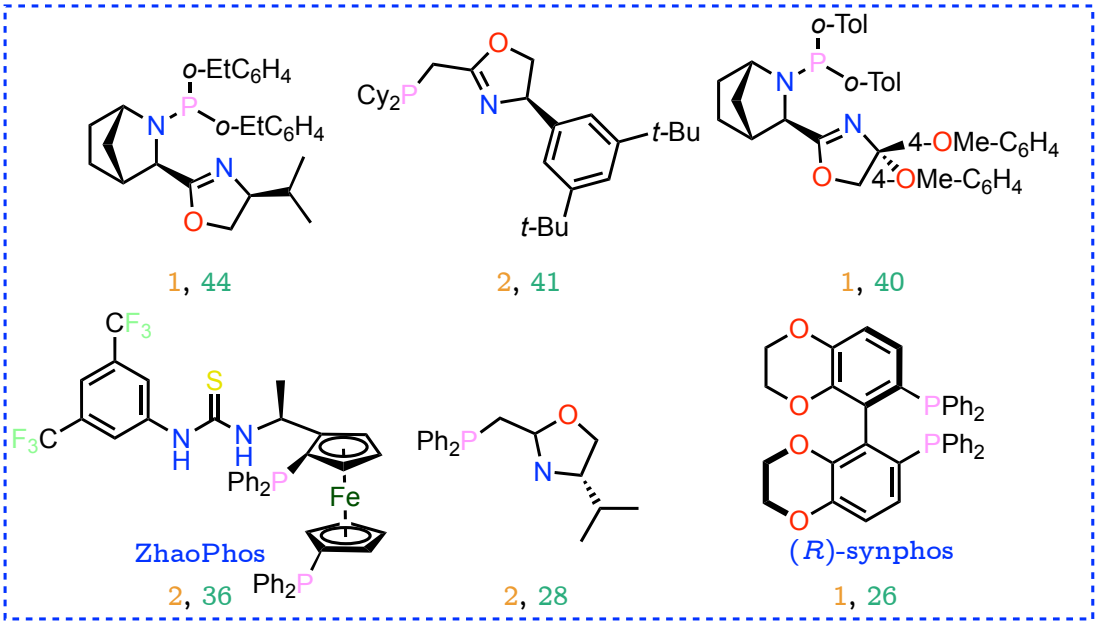


Figure 2.5: Chemical structures of the six most frequently reported ligands in the dataset, ranked by total reaction count. Reaction counts are shown in green, and publication counts in orange.

The distinction between reaction counts and publication counts for both substrates and ligands underpins the trade-off between breadth of application and depth of study, which has important implications for model training and generalisability. When the dataset contains many reactions of the same ligand or substrate from a single publication, the model may overfit to the specific reaction conditions of that study. Structures observed across multiple publications reflect broadly applicable reaction conditions. Therefore, reaction heavy datasets provide depth but risk poor generalisability, while publication diverse datasets offer breadth but less data per scaffold can reduce structure specific predictive ability, but increase overall generalisability. Striking the right balance between depth and breadth can be challenging. In this project, the choice was made to maximise the total number of reactions available for model training by including all reactions reported.

To examine the distribution of ligand types in the dataset, ligands were grouped into categories according to their scaffold type. This categorisation was used as a feature in the ML models (**Table 2.2**). Representative examples of the most prevalent ligands for each scaffold type are shown in **Figure 2.6**.

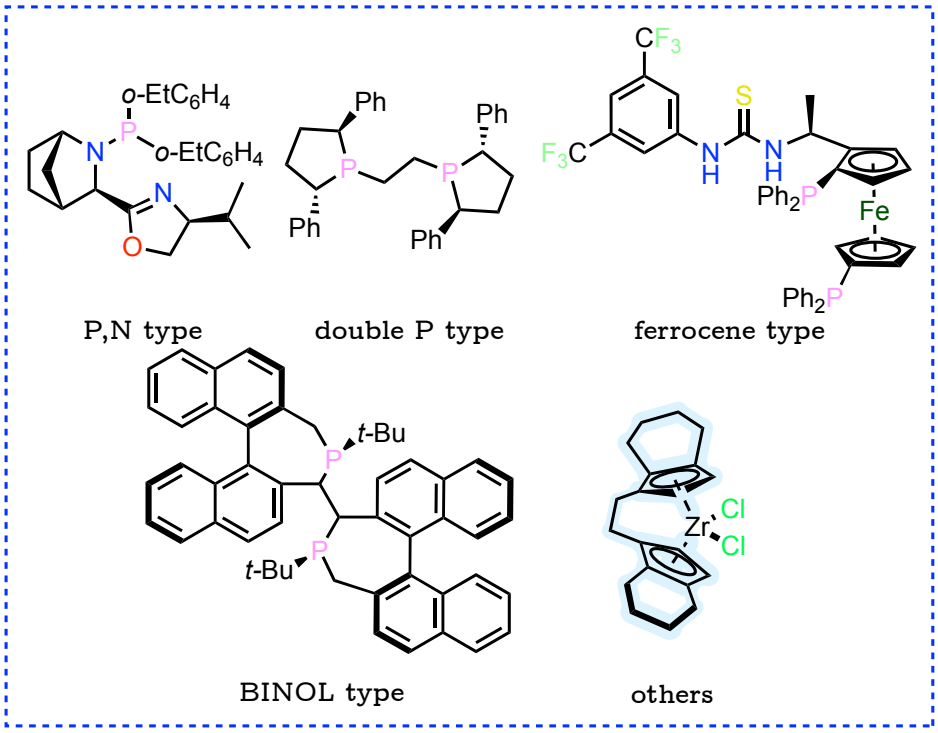


Figure 2.6: Examples of top-ranked ligands by reaction number for each scaffold type.

The relative distribution of catalysts is shown in **Figure 2.7**, categorised by metal centre and ligand scaffold type, based on the number of reactions reported for each. The left panel shows that iridium (44.2%) and rhodium (34.0%) dominate the metal distribution, together accounting for nearly 80% of all reactions. Ruthenium follows at 10.9%, with other metals such as cobalt, nickel, palladium, and zirconium contributing less than 6% each.

The right panel shows the scaffold type distribution, in which *P,N* type (39.0%) and double P type (38.7%) are the most prevalent, making up nearly 78% of all cases. Ferrocene type (13.2%) and BINOL type (6.9%) scaffold appear less frequently, and any other types are rare (2.1%).

The heatmap in **Figure 2.7** reveals a strong association between metal and scaffold type, indicating that certain metals are predominantly paired with specific ligand structures. Rhodium is most frequently associated with bisphosphine ligands (double P type), while iridium shows a clear preference for *P,N*-type scaffolds. This reflects the discussion of the landscape of this reaction in Chapter 1. Other metals that are less prevalent in the dataset, display more diverse scaffolds, but remain dominated

by one or two categories. Additionally, several metal-scaffold combinations remain completely unexplored. This skew has implications for ML as the model will be heavily influenced by the selection bias by having reduced predictive capability for the underrepresented combinations. However, the less explored regions highlight areas that could offer unique opportunities for future catalyst development.

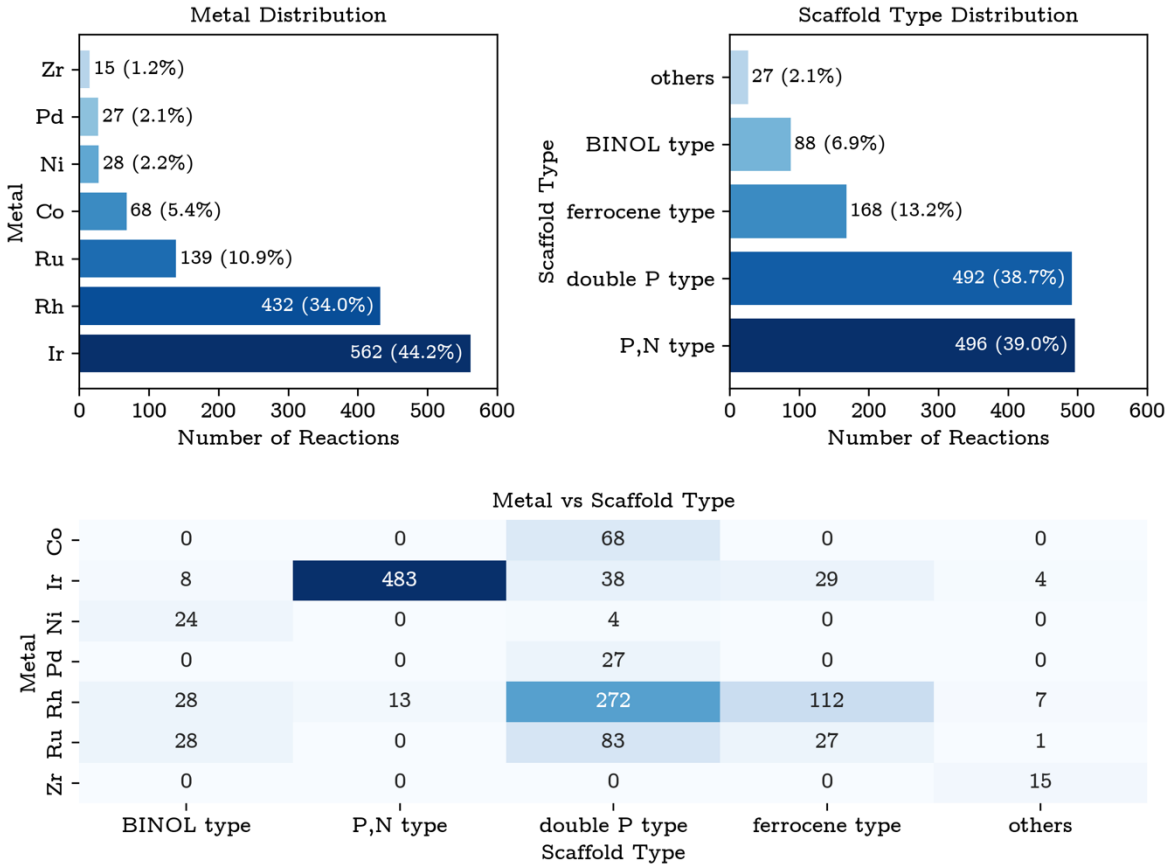


Figure 2.7: Metal and scaffold type distributions (top) and their relationship (bottom). The top left panel shows the number and percentage of reactions for each metal, while the top tight panel shows the distribution of scaffold types, with less common types grouped as 'others'. The bottom panel presents a heatmap of reaction counts for each metal-scaffold combination.

The distribution of solvents and additives in the dataset (**Figure 2.8**) are also highly skewed, with a small number of entries dominating the reported reactions. DCM and MeOH are used as solvents in most cases, reflecting their widespread use in homogenous catalysis. Over 80% of the reactions were conducted with no additive, but tetrafluoroboric acid (HBF_4), triethylamine (TEA), trifluoroacetic acid (TFA) and potassium tert-butoxide (*t*BuOK) are the most frequent and have been used

successfully in the literature. Many solvents and additives occur very rarely, with entries representing less than 3% of solvents and 1% of additives grouped as 'others' in this figure. This could again bias the ML model towards the more well-represented combinations and be unable to successfully predict the reactions of these unexplored areas.

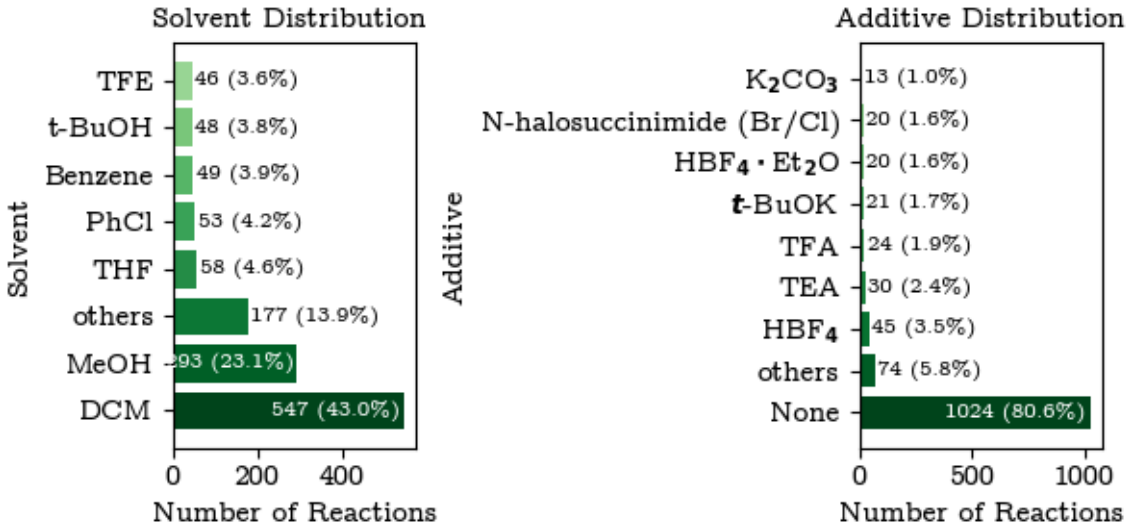


Figure 2.8: Distribution of solvents (left) and additives (right) used in the dataset. Rare entries were grouped into an 'others' category (solvents: <3% of reactions; additives: <1%). Labels show the number of reactions and corresponding percentage for each entry.

Figure 2.9 shows the temperature and pressure distributions for the dataset. Most reactions (665) were conducted at room temperature (20 – 25 °C), likely to preserve enantioselectivity, while the remainder were carried out at elevated temperatures up to 104 °C. Reaction pressures exhibited greater variability, with the kernel density estimate (KDE) revealing multiple peaks, most prominently around 50 bar H_2, followed by peaks near 10, 30, and 100 bar H_2.

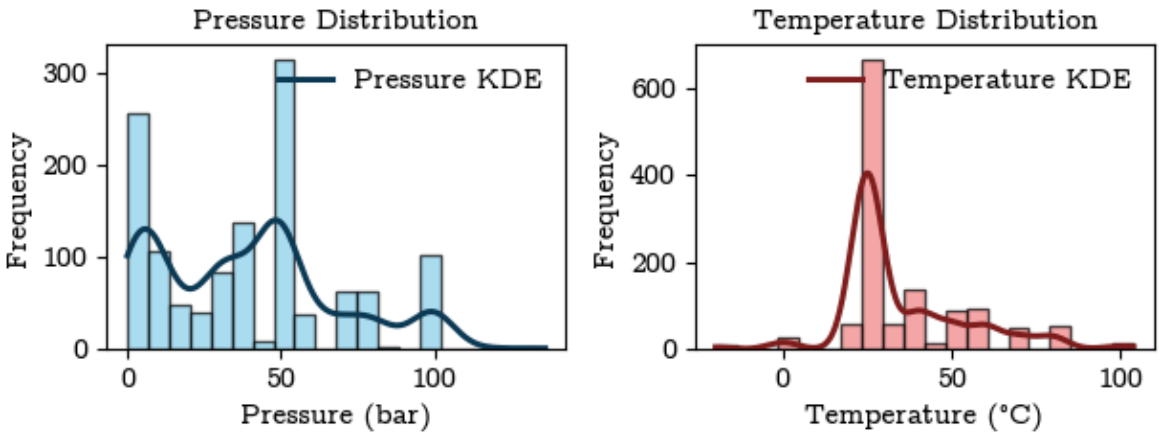


Figure 2.9: Pressure (left) and temperature (right) distributions in the dataset, shown as histograms with overlaid KDE. Histograms display the frequency of reported conditions, while KDE curves highlight the underlying distribution trends.

Figures 2.7, **2.8**, and **2.9** highlighted the skew caused by selection bias inherent in datasets derived from literature sources, due to the human tendency to rely on established synthetic routes. Such datasets also reflect a strong reporting bias, with unsuccessful or low-performing outcomes rarely disclosed. This bias is seen in this dataset, evident in the distribution of *ee* (**Figure 2.10**). 648 reactions achieve over 90% *ee* while low and moderate *ee* values are comparatively rare. This presents a challenge for developing balanced ML models capable of accurately predicting *ee*, specifically in the low-mid range.

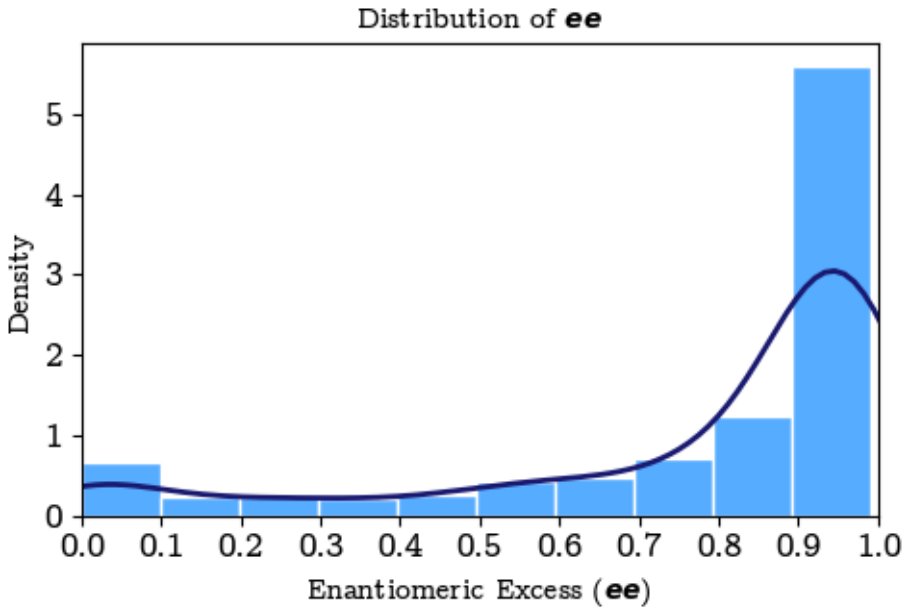


Figure 2.10: Distribution of ee values in the dataset, shown as a histogram with an overlaid KDE

2.3.2 Regression Machine Learning Models

To predict the *ee* of AH reaction of TSC=C, regression ML models were trained on 80% of the dataset using three sets of input features (2D descriptors, 3D descriptors, and 2D+3D descriptors) with the true *ee* values as the target output. The optimum combination of hyperparameters found for each model algorithm, and therefore used for model training, are displayed in the Appendix. Following training, the ML models were used to predict the *ee* of the remaining 20% of the dataset, and the resulting MAE between the ML predictions and the true values and SD are shown in **Figure 2.11**. An MAE of under 10% *ee* was set as the performance target for this project, as this level of accuracy was thought to be sufficient to guide the early stages of screening in an optimisation campaign.

Figure 2.11 shows that model class influenced test MAE more strongly than the descriptor set, with tree-based ensembles outperforming linear models across the different descriptor sets. The best-performing model was ETR(2D+3D) with a test MAE of 9.8 ± 1.2% *ee*, the only model below the 10% *ee* MAE target. Relative to the same algorithm using 2D descriptors only (ETR(2D): 11.9 ± 1.3% *ee*), the combined input model achieved an absolute improvement of 2.1% *ee* (a 17.6% reduction),

indicating that 2D and 3D descriptors provided a complementary pattern that this model could exploit.

A similar observation was seen for GBR, where GBR(2D+3D) outperformed both GBR(2D) and GBR(3D). In addition, models based solely on 2D features generally outperformed those based solely on 3D features across these ensembles. By contrast, adding 3D descriptors did not improve the linear models (LLR, RR) or RFR. For example, LLR(2D) and LLR (2D+3D) both showed an MAE of 15.4 ± 1.3% *ee*, each outperforming LLR(3D). This lack of improvement with the addition of 3D descriptors for linear models likely reflect their limited capacity to capture increased variance in the features. Despite the algorithms similarity to ETR/GBR, RFR performed less well, likely due to its different randomisation strategy relative to ETR.

SVR was evaluated on the 3D subset only due to its higher training time when high-dimensional ECFP features were included. However, although training time increased with feature count, ECFPs can be derived directly from SMILES and therefore avoid QC calculations. Therefore, where comparable accuracy is attainable without QC descriptors, the models trained on 2D features alone can offer a computationally faster and more scalable approach.

Error bars (±1 SD over repeated 80/20 splits) indicate moderate sensitivity to the train/test partition, typically 0.6–1.3% *ee* MAE. Several mid-tier models (MAE ≈ 12–15% *ee*) show overlapping intervals, as such rank differences among them may not be substantially meaningful. Additionally, the ETR(2D+3D) SD straddles the 10% MAE *ee* threshold.

For predictive use, ETR(2D+3D) was the best-performing in terms of predictive ability, with ETR(2D) second best, which may be advantageous in settings where omitting 3D/QC features and thus reducing computational cost is beneficial.

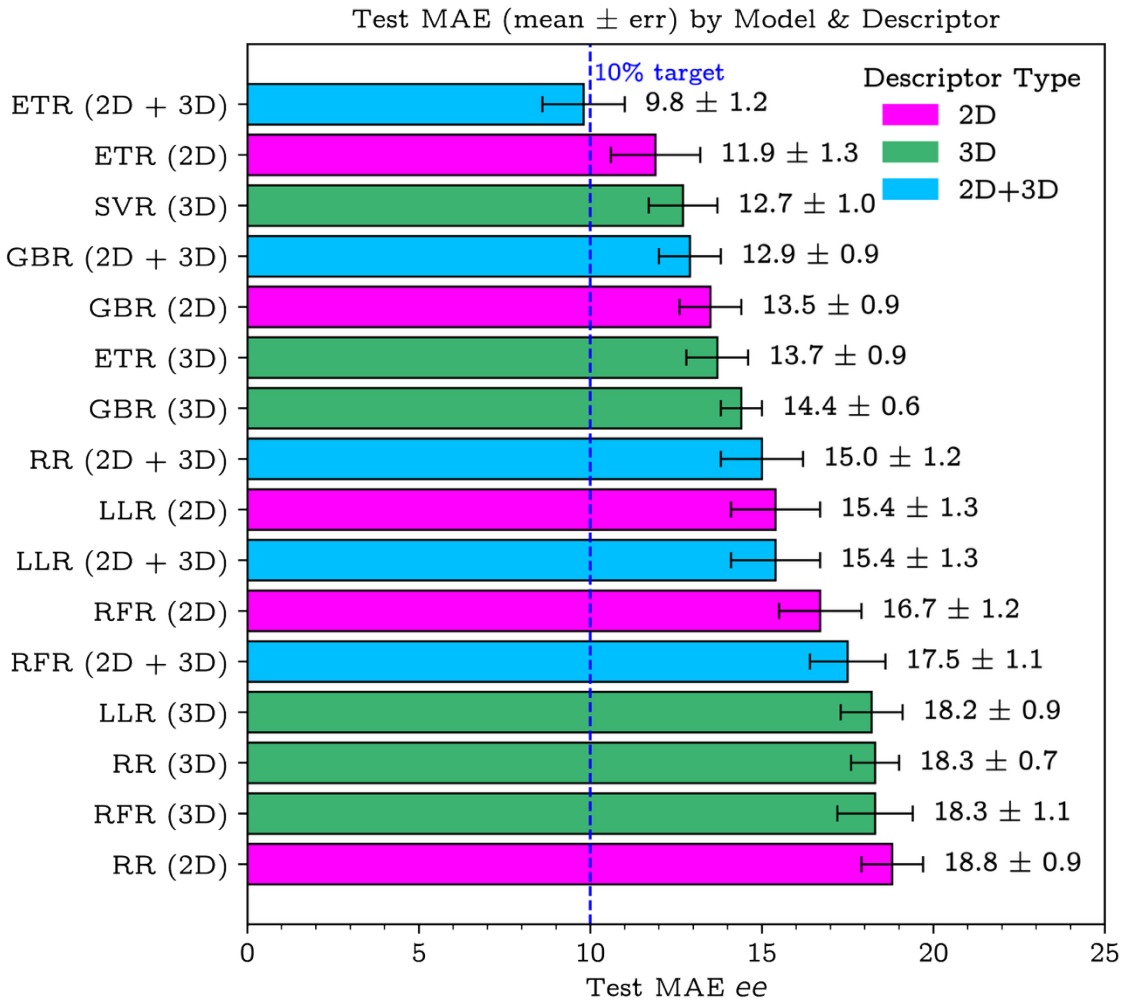


Figure 2.11: Test-set performance of all regression models across descriptor sets. Reported in percentage points of *ee* (% *ee*). Bars show MAE with ± SD, computed over different train-test splits, evaluation is always on the held-out portion. Models are ordered by increasing MAE, and colours indicate descriptor type (purple = 2D, green = 3D, blue = 2D+3D). The vertical line at 10% marks the target MAE threshold.

Learning curves were generated for all fitted models (see Appendix). **Figure 2.12** shows the curves for the best-performing model ETR(2D+3D), and ETR(2D) shown for comparison. In both cases, the CV test MAE decreases as the training set grows and the SD bands narrow, indicating that the models are learning a genuine signal and predictions become more stable with additional data. The training MAE remains below the test MAE, and the two curves have not converged at the largest sample size, consistent with moderate overfitting. Notably, the test curve continues to trend downward at the maximum training size, suggesting that additional data would likely improve the model's performance.

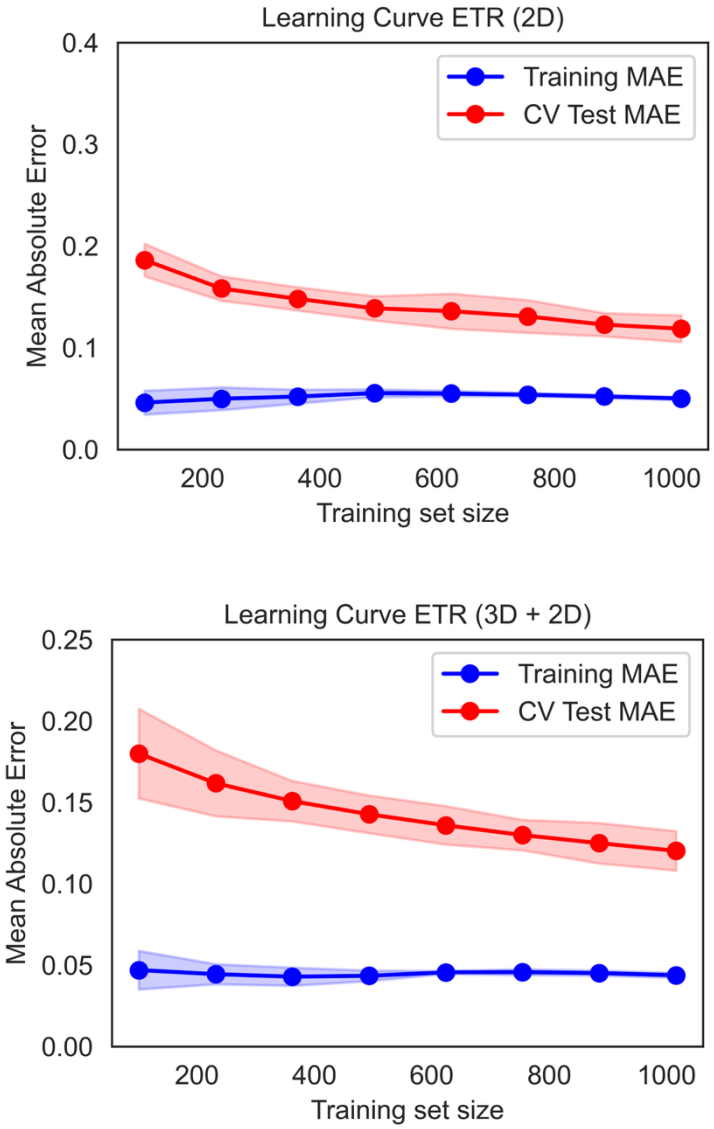


Figure 2.12: Learning curves for the ETR(2D) and ETR(3D+2D) models, where *ee* is represented as a fractional value. An MAE of 0.10 corresponds to 10% *ee*. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

The learning curves for the GBR(2D) and the GBR(2D+3D) models are shown in **Figure 2.13** and show a similar pattern as the ETR models.

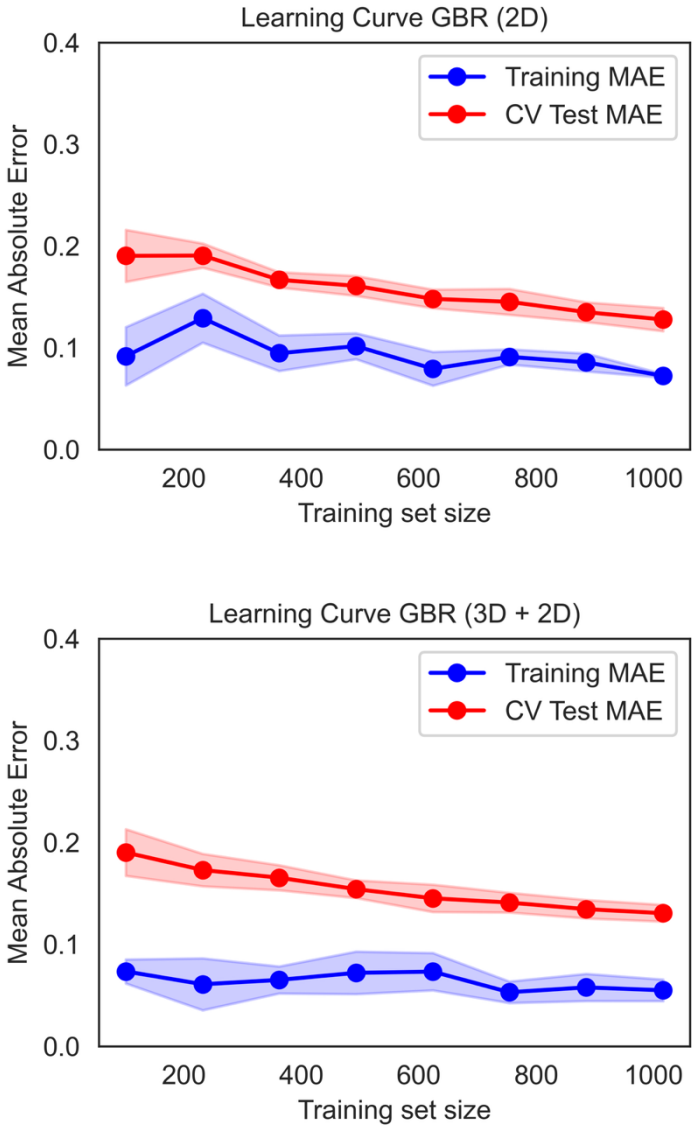


Figure 2.13: Learning curves for the GBR(2D) and GBR(3D+2D) models, where *ee* is represented as a fractional value. An MAE of 0.10 corresponds to 10% *ee*. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

Figure 2.14 shows the RFR(2D) learning curve contrasts with the patterns observed for ETR and GBR. The training and CV test MAE curves nearly converge, and their uncertainty bands overlap, indicating no overfitting. However, this observation combined with the higher MAEs seen for RFR relative to the best models, suggest that this model algorithm may not capture the full complexity of the dataset.

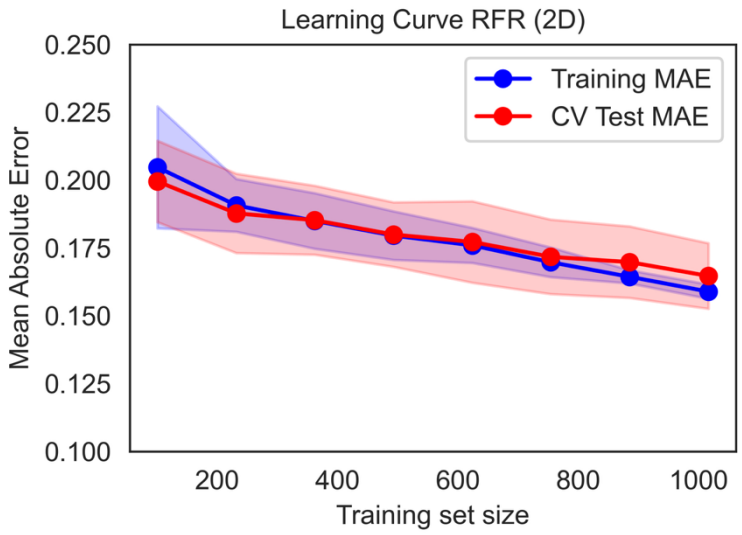


Figure 2.14: Learning curves for the RFR(2D) model, where *ee* is represented as a fractional value. An MAE of 0.10 corresponds to 10% *ee*. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

To evaluate whether the uneven distribution of *ee* in the dataset affects predictive performance, the predicted vs. true *ee* and residual histograms (predicted − true) were generated for each model (see appendix). The plots for the ETR models are shown in **Figure 2.15**. The residual histograms are centred at zero, indicating that most predictions are close to the true values. However, the scatter plots show that the low *ee* cases tend to be overpredicted, whereas the high *ee* region is generally well captured. This pattern is seen across all input features and is consistent with the unbalanced *ee* distribution in the dataset. This heavily suggests that additional data in the low *ee* regime could reduce this bias.

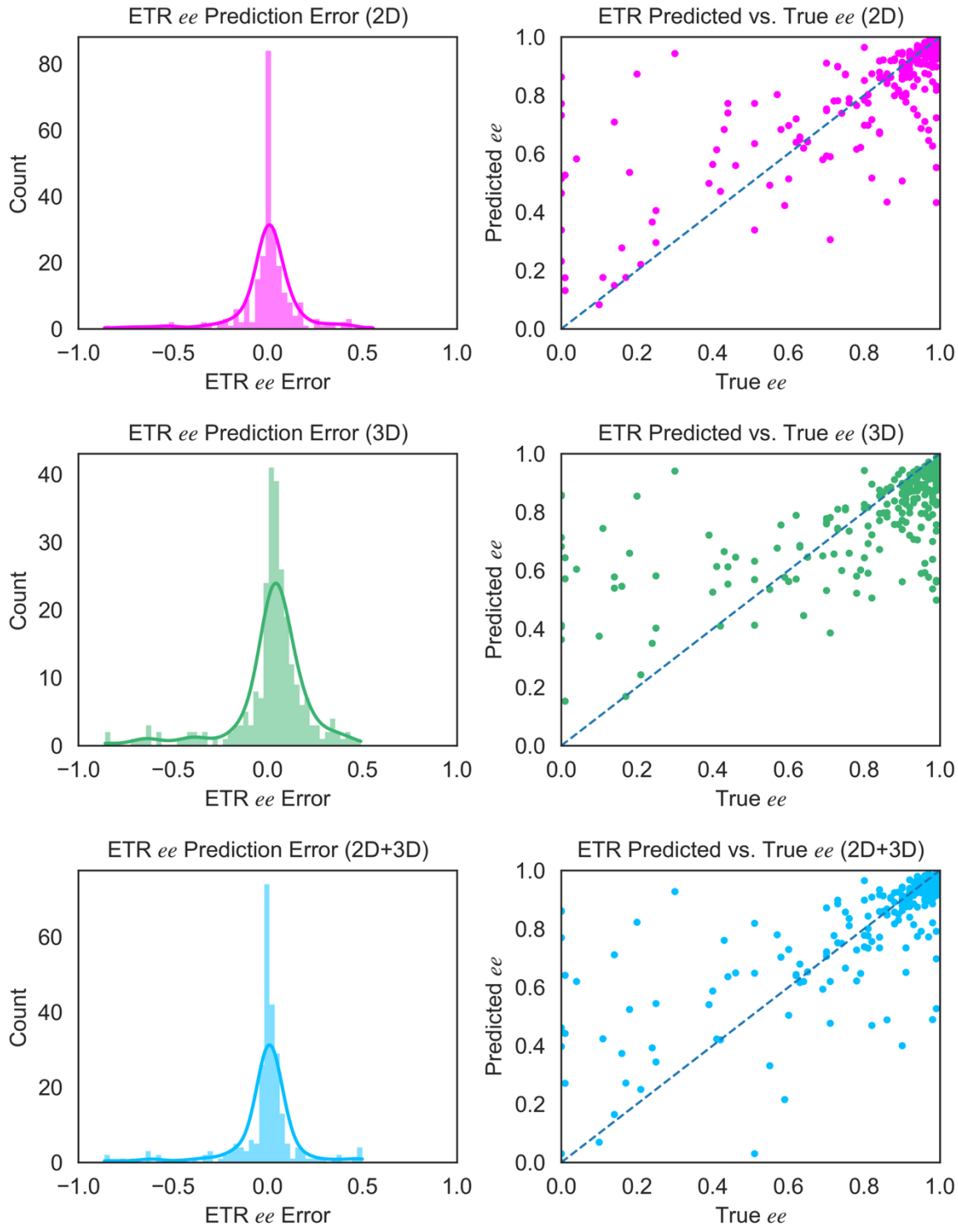


Figure 2.15: : Performance of the ETR model for *ee* prediction across the different sets of input features. (Left) Distribution of prediction errors on the test set, with the KDE curve overlaid. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

The permutation feature importances were calculated for each of the fitted models (see Appendix) and the plotted permutation importances for ETR(2D+3D), GBR(2D+3D)

and SVR(3D) are shown in **Figures 2.16**, **2.17** and **2.18**. The results indicate that the prediction is influenced by a complex combination of steric, electronic factors as well as being heavily influenced by reaction conditions such as solvent and pressure. The SVR(3D) model highlighted steric descriptors as particularly important with a sterimol parameter around the metal (c_sterimol_metal_L (L1)) showing a higher importance than other features. ETR(2D+3D) and GBR(2D+3D) distribute importance more evenly across the feature set. The consistently high importance of the solvent and pressure used across models emphasises the importance of the reaction conditions to enantioselectivity. Interestingly, the ETR(2D+3D) model is governed more by chemical structure, rather than electronic and steric parameters, indicated by the ECFP features found to be most important.

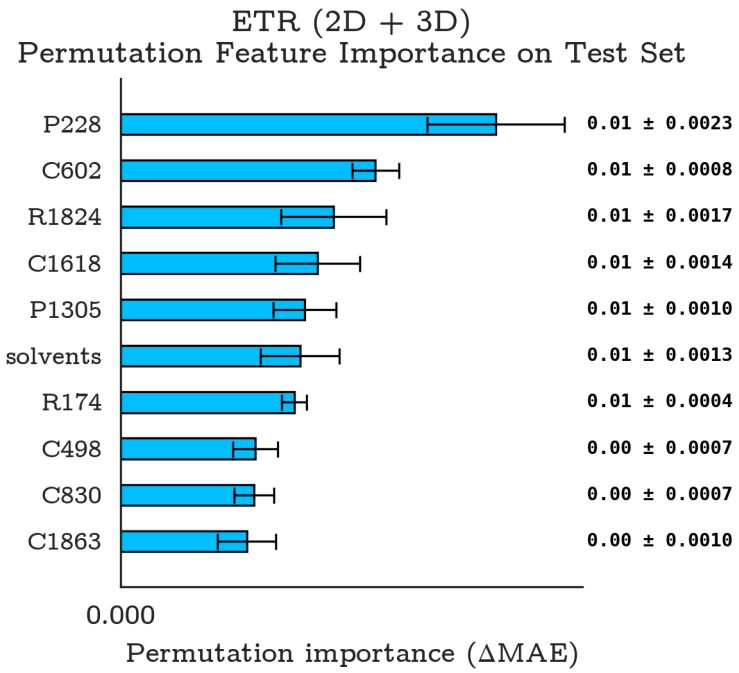


Figure 2.16: Permutation feature importance for the ETR model trained on the combined 2D + 3D descriptor set. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

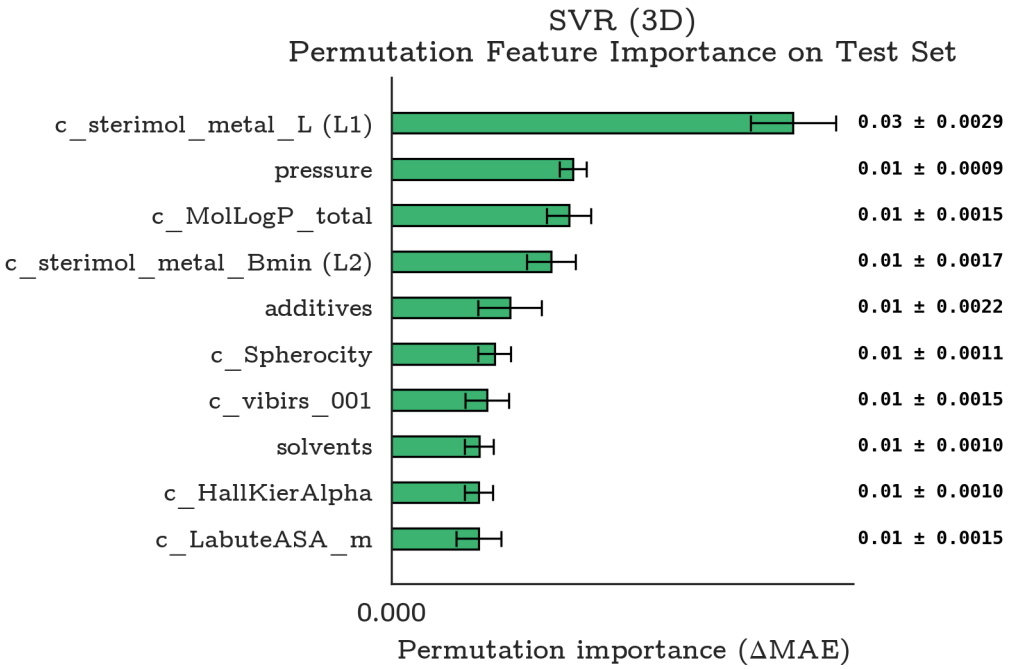


Figure 2.17: Permutation feature importance for the SVR model trained on the 3D descriptor set. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

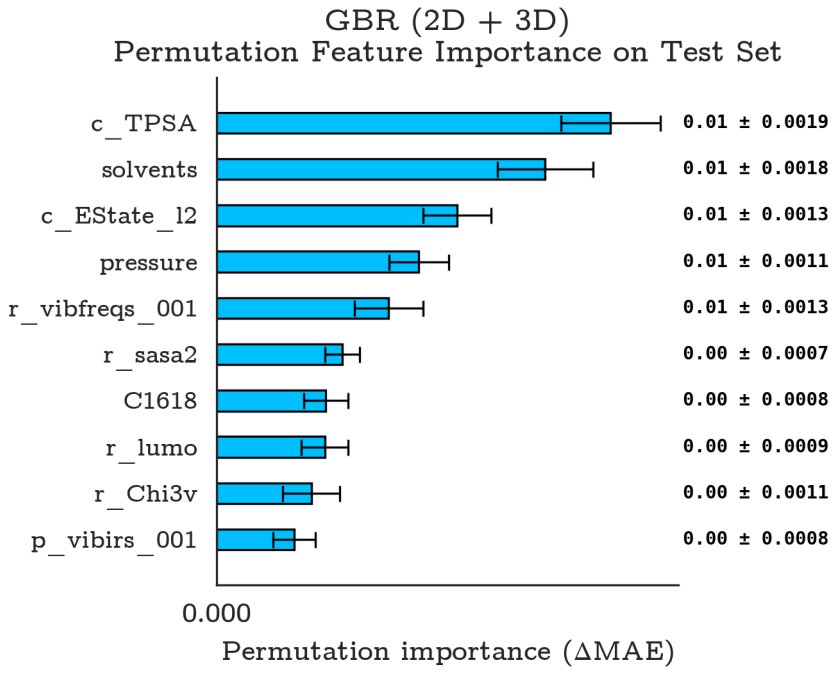


Figure 2.18: Permutation feature importance for the GBR model trained on the combined 2D + 3D descriptor set. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

2.3.3 Classification Machine Learning Models

Regression models showed reduced predictive abilities for low–mid *ee* values, where limited data was available. The problem was therefore reformulated as a classification task by binning *ee* into two categories, high (≥ 80% *ee*) and low (< 80% *ee*), and models were trained to predict class membership rather than exact values. This was expected to improve predictive ability relative to the direct regression models by increasing data density in this low *ee* region. Additionally, although not giving a prediction of the exact *ee* value, resulting models built on this prediction could determine whether a reaction would be selective enough for consideration at the early stages of an optimisation campaign. The classes are shown in **Figure 2.19,** after categorisation below and equal to or above 80% *ee*. The 1271 reactions were split resulting in 863 (67.9%) reactions in the high class and 408 (32.1%) in the low class. This reinforces the class imbalance that exists in the dataset, with the high *ee* class having the majority of reactions.

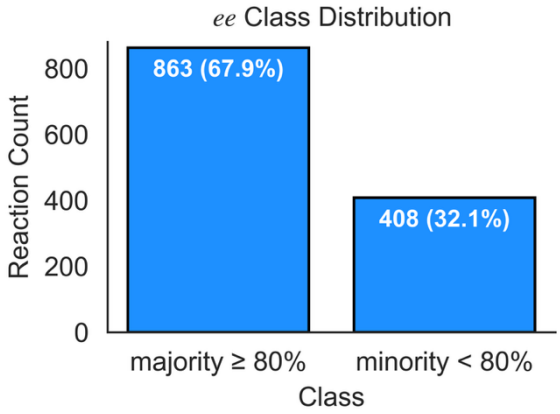


Figure 2.19: Distribution of ee classes after binarization at the 80% threshold.

The accuracies obtained for the percentage of correct predictions and their relative SD are shown in **Figure 2.20**. Across all algorithms and descriptor sets the test set accuracy fell into 76 - 81%. The highest accuracy was given by the GBC model with 3D input features, resulting in an accuracy of 81.4 ± 1.1%. Given the class imbalance (high *ee* class being 67.9%), all models exceeded this majority class baseline by 8-14%, indicating meaningful predictions were produced. Error bars between the models and

input features overlap, suggesting a limited separation of performance. A small advantage was observed for 3D descriptors over 2D and combining 2D+3D feature sets.

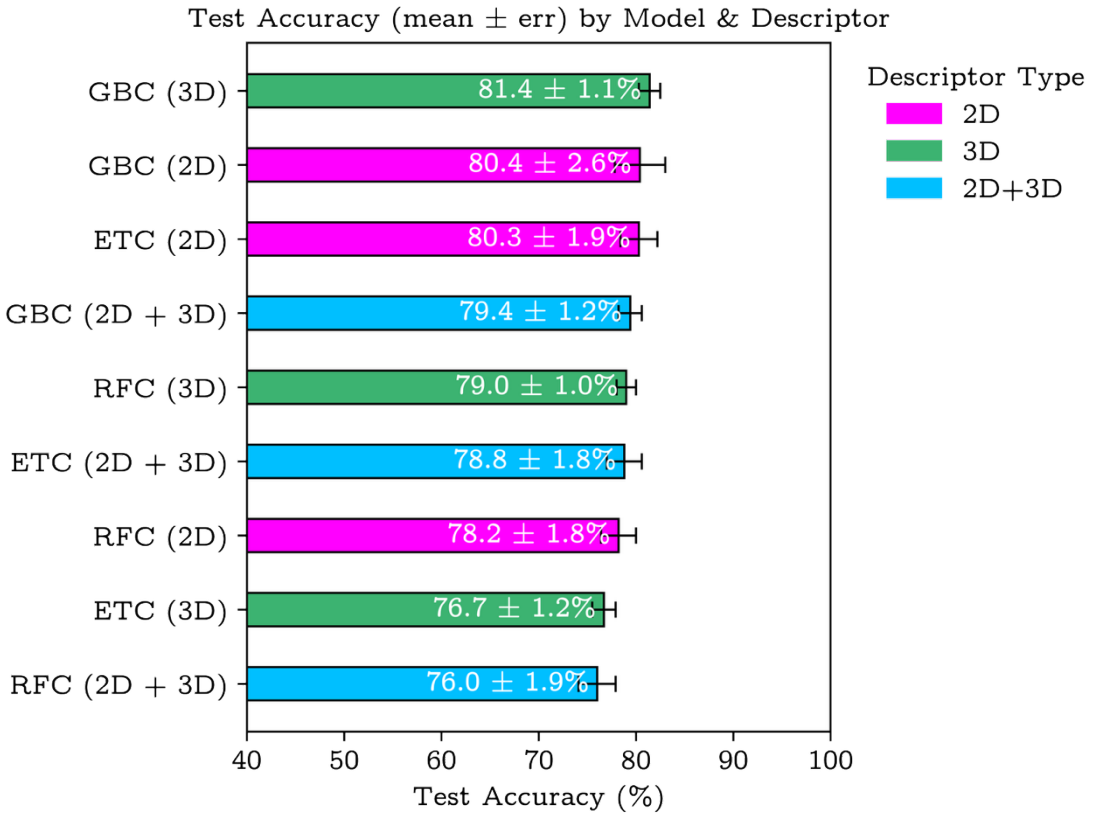


Figure 2.20: Test accuracy (mean ± SD across train-test splits) for *ee* classification (high: ≥80% *ee*; low: <80% *ee*) by algorithm and descriptor set, reported as a percentage. Bars are coloured by descriptor type (purple: 2D; green: 3D; blue: 2D+3D); error bars show ±1 SD.

To examine the breakdown of the accuracies for each of the classes, confusion matrices were plotted for all fitted models and are shown in **Figure 2.21**, **2.22** and **2.23**. The matrices reveal a consistent asymmetry in class performance. Across all algorithms and descriptor sets, the class prediction for the high *ee* class (80-94% accuracy) was stronger than for the low *ee* class (41-72% accuracy). Consequently, the dominant error in the models is false positive, where low-*ee* reactions were predicted as high, reflecting the 67.9% majority of high-*ee* samples in the data.

For 2D descriptors (**Figure 2.21**), GBC achieves the highest overall accuracy for this set of descriptors and shows strong high-class prediction (92% accuracy) with moderate low-class prediction (62% accuracy). ETC attains the highest high-class accuracy (94%) but at the expense of lower low-class accuracy (56%). RFC provides

the best low-class prediction in the 2D set (72% accuracy), resulting in a balanced accuracy of 78%, despite lower overall mean accuracy than GBC.

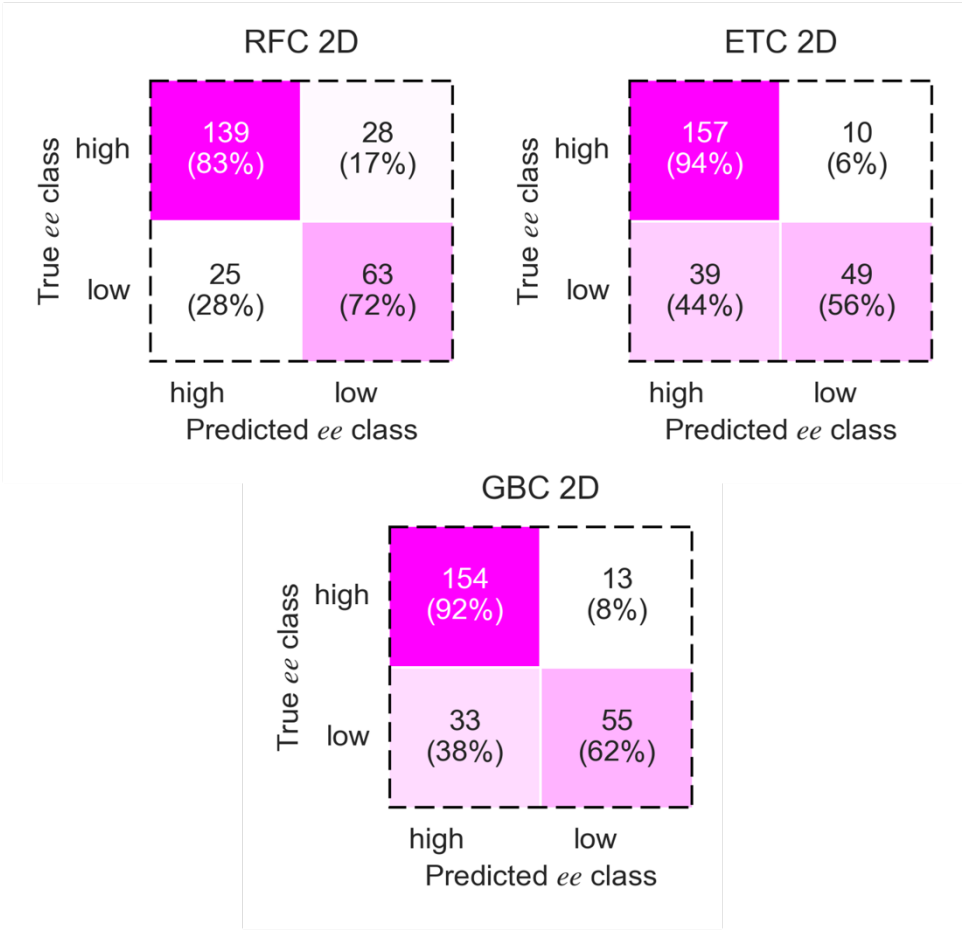


Figure 2.21: Confusion matrices for 2D descriptor classifiers, RFC, ETC and GBC on the classification task (high ≥ 80% ee, low <80% ee). Cell entries are counts with per class accuracy percentages given.

For 3D descriptors (**Figure 2.22**), ETC improves low-class accuracy to 68% while maintaining 80% accuracy on the high class, RFC drops to 47% low-class accuracy despite 92% accuracy on the high class. The highest performing GBC model balances these at 89% (high-class accuracy) and 58% (low-class accuracy).

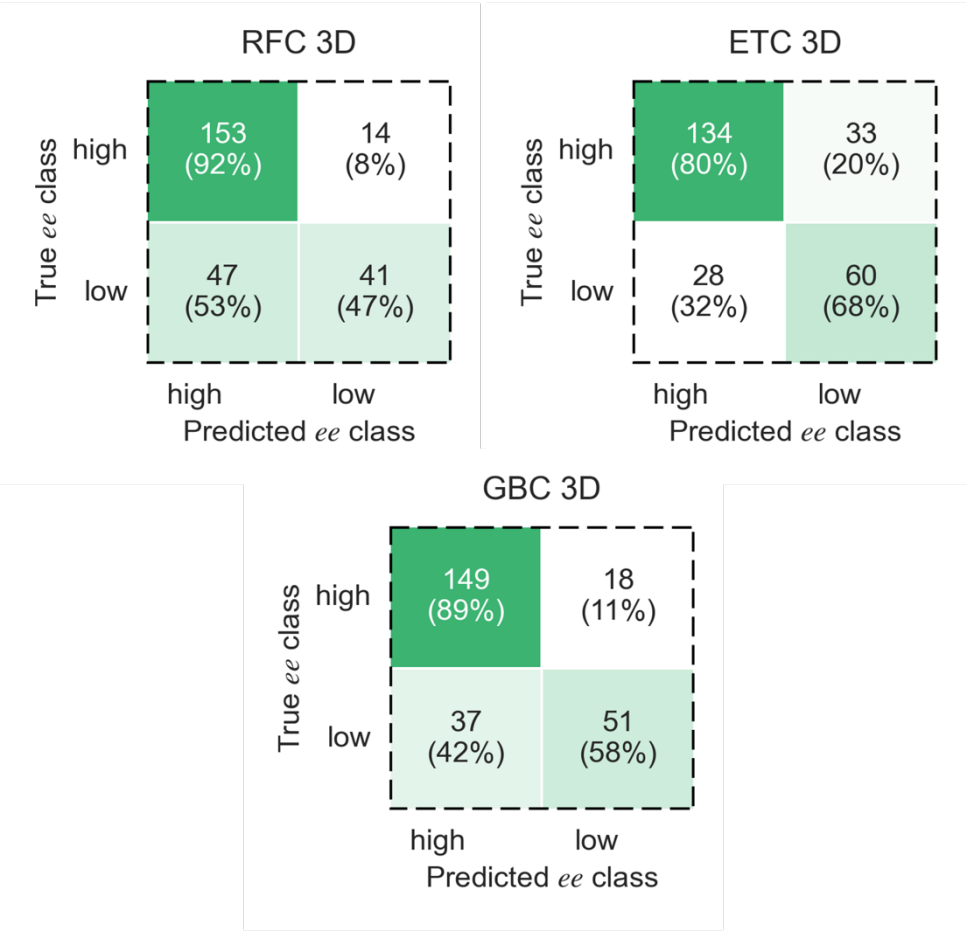


Figure 2.22: Confusion matrices for 3D descriptor classifiers, RFC, ETC and GBC on the classification task (high ≥ 80% ee, low <80% ee). Cell entries are counts with per class accuracy percentages given.

For combined 2D+3D descriptors (**Figure 2.23**), performances were similar to the singular input descriptor models, indicating there was limited complementary patterns found in this dataset by these models. However, GBC retains very high high-class accuracy (92%) but with poorer low-class accuracy (53%). Whereas RFC improve low-class accuracy to 65%.

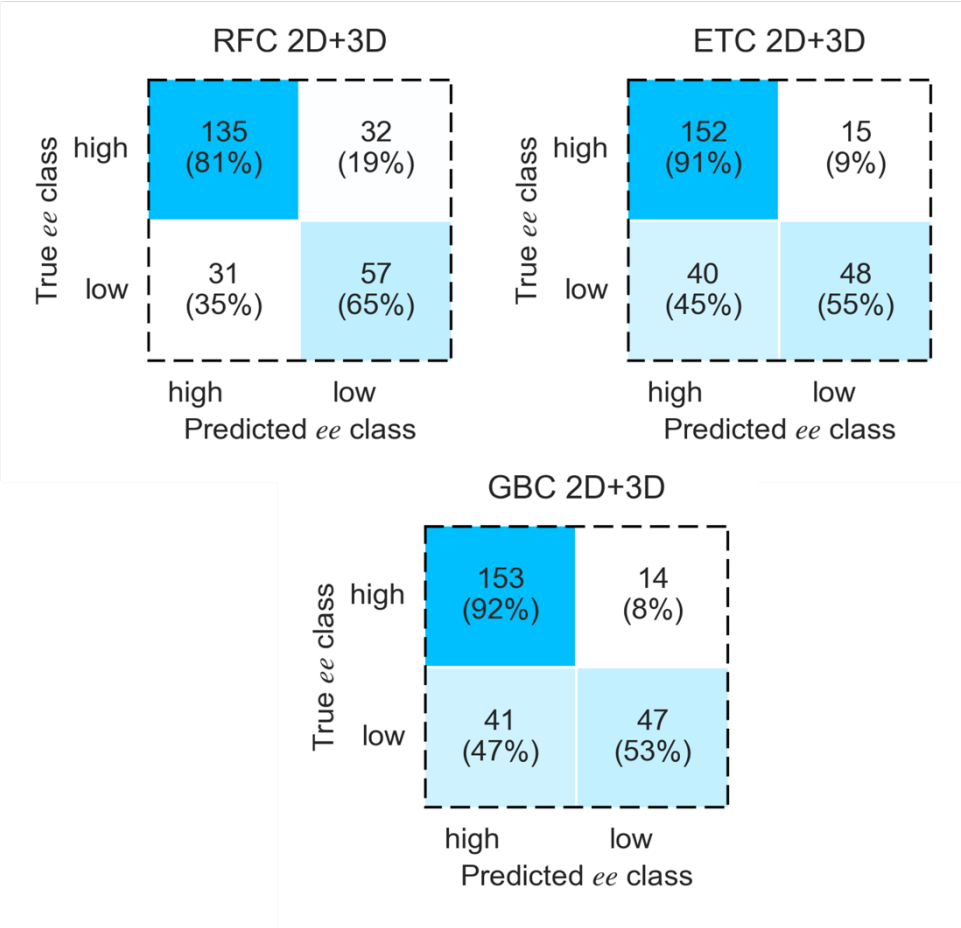


Figure 2.23: Confusion matrices for 2D+3D descriptor classifiers, RFC, ETC and GBC on the classification task (high ≥ 80% ee, low <80% ee). Cell entries are counts with per class accuracy percentages given.

Learning curves were fit for all fitted classification models and can be seen in the Appendix. The curves for GBC(3D), ETC(3D) and RFC(2D) are shown in **Figure 2.24**, **2.25** and **2.26**. Across models, the test accuracy increases with training size, confirming learnable patterns in the features. The training-test gaps are modest, indicating some variability and overfitting. However, this narrows with training sample size, consistent with reducing the dataset noise.

The highest-accuracy model GBC(3D) (**Figure 2.24**), shows a persistent gap between the train and test accuracies, suggesting this model is still variance limited at the available sample size. This aligns with its highest performing accuracy (81%) and the confusion matrix pattern, showing much better accuracy for the high *ee* class.

Additional data would be expected to close this gap and improve distinction of the low *ee* class.

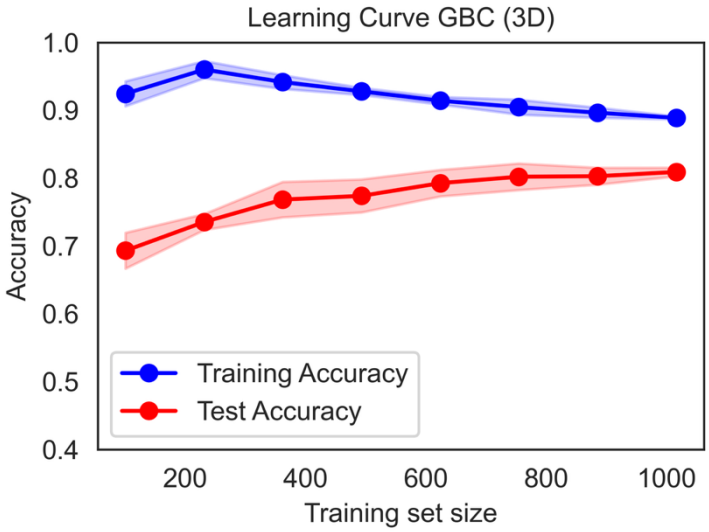


Figure 2.24: Learning curves for the GBC(3D) model. Blue markers show the CV training accuracy and the red markers show the CV test accuracy, reported as fractional accuracy (0-1 scale, where 1.0 corresponds to 100% accuracy). The shaded bands denote the SD.

The ETC(3D) model (**Figure 2.25**), shows a smaller more stable train-test gap and a steady rise in both curves. This is characteristic of lower model variance compared to GBC(3D). The corresponding confusion matrix showed improved low *ee* accuracy to the other 3D models, indicating this model captures discriminative patterns for the low *ee* class without overfitting. Again, the upward slope at the largest dataset size suggests that additional data should result in improved performance.

The RFC(2D) (**Figure 2.26**), shows a gap comparable to ETC. At the smaller sample size, the train-test gap is larger, but this narrows considerably with more data. Despite slightly lower overall accuracy than the GBC(3D), the confusion matrix revealed a relative strong prediction of the low *ee* class (highest low class accuracy among 2D models). The continued upward trend of both curves indicates that again this model remained data limited with this set of input features.

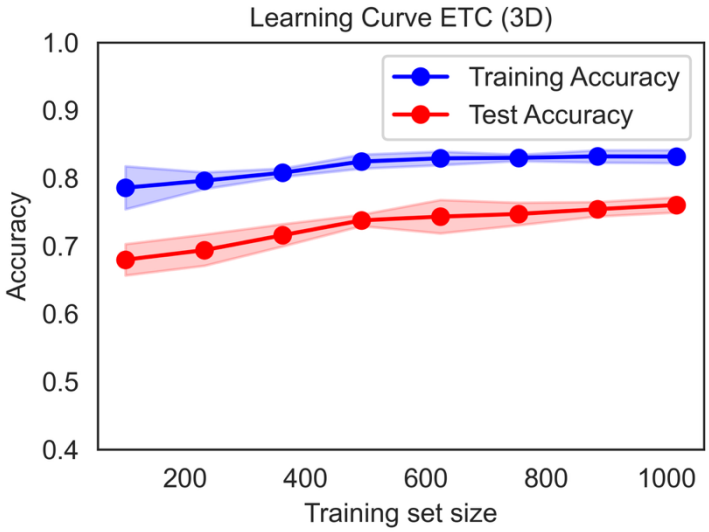


Figure 2.25: Learning curves for the ETC(3D) model. Blue markers show the CV training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

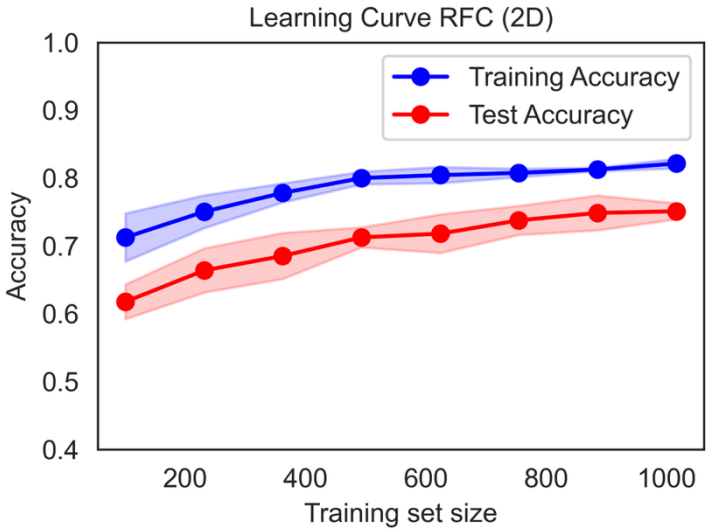


Figure 2.26: Learning curves for the RFC(3D) model. Blue markers show the CV training accuracy and the red markers show the CV test accuracy, reported as fractional accuracy (0-1 scale, where 1.0 corresponds to 100% accuracy).The shaded bands denote the SD.

The permutation feature importances were fitted for all classification models and are shown in the Appendix. The feature importances for GBC(3D) and ETC(3D) are shown in **Figure 2.27** and **2.28**. Overall, the feature importances on the test set indicated multiple features drove predictions, rather than a single dominant feature.

For GBC(3D) (**Figure 2.27**), the most influential features combined the local 3D sterics around the metal, such as the sterimol lengths, with the polarity (TPSA) and electronic descriptors (partial charges and frontier orbital energies) from the reactants, products and catalysts.

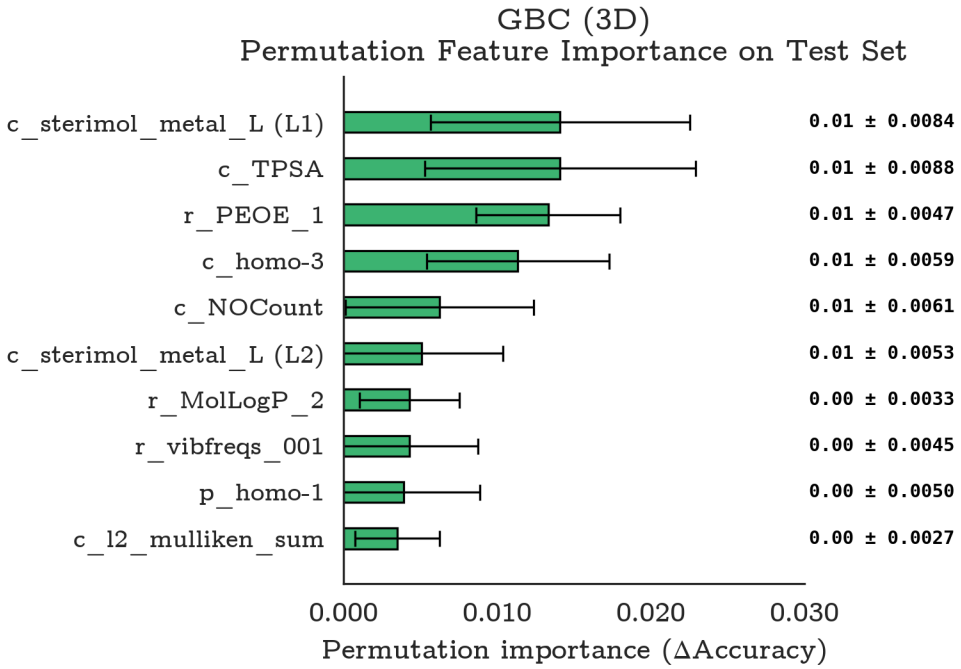


Figure 2.27: Permutation feature importance for the GBC model trained on the 3D descriptor set. Bars show the mean increase in test accuracy when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

For ETC(3D) (**Figure 2.28**), importance shifted towards heteroatom counts (NO count, H-bond acceptors), SASA on the metal catalysts, alongside the experimental reaction conditions (additives, pressure, catalyst loading). The overlap in heteroatom and polarity features across the two models suggests a shared pattern found from these input features. The GBCs emphasis on steric/electronics and ETCs emphasis on counts and conditions could indicate their confusion matrix behaviour of the higher overall accuracy for GBC(3D) and stronger low *ee* accuracy for ETC(3D).

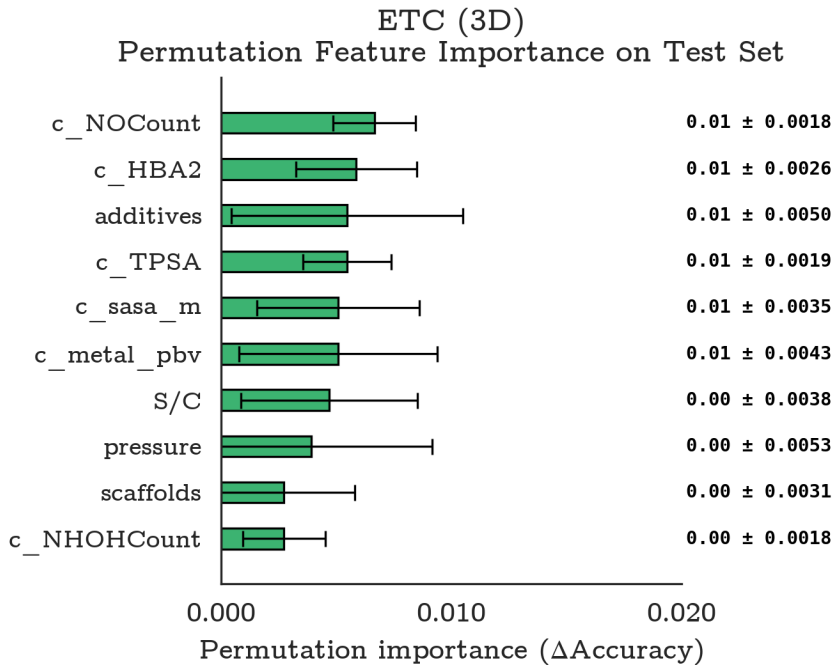


Figure 2.28: Permutation feature importance for the GBC model trained on the 3D descriptor set. Bars show the mean increase in test accuracy when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

2.3.4 Synthetic Minority Over-sampling of Classification Models

To counteract the class imbalance that exists in the dataset, the minority class (<80% *ee*) of the training set was augmented with synthetic samples using SMOTE. This approach was better suited to classification, where class frequencies could be equalised, than to the regression models. After implementing SMOTE on the training set, the classes were balanced from containing 326 low *ee* class reactions and 690 high *ee* class reactions, to both classes containing 690 data points. This dataset was used to train classifications models on the GBC and ETC algorithms, which showed good predictive performance previously, but struggled with predictions for the low *ee* class (minority class). Due to the generation method of SMOTE, synthetic data was only added on the 3D feature subset due to their greater interpretability, compared to the synthetic generation of fingerprints. The accuracy and SDs of the fitted models and their confusion matrices are shown in **Figure 2.29.**

For GBC(3D), the test-set accuracy decreased slightly from 81.4 ± 1.1% (no SMOTE) to 80.7 ± 1.5% with SMOTE. The confusion matrix shows the low class accuracies rising from 58% to 76%, while high class accuracy decreased (89% to 85%).

The main failure observed without the inclusion of SMOTE data, misclassifying low *ee* as high *ee,* was reduced by 18%, but remained the biggest source of error for this model.

For ETC(3D), overall accuracy improved (76.7 ± 1.2% to 78.6 ± 1.6%). Low class predictive accuracy increased from 68% to 77%, accompanied by a small increase in the high class accuracy 80% to 81%.

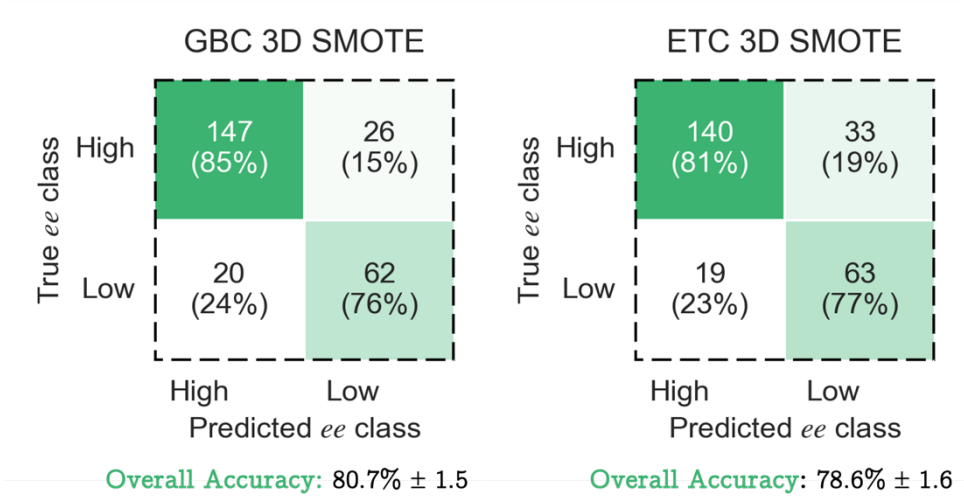


Figure 2.29: Confusion matrices for classifiers with added SMOTE data, ETC(3D) and GBC(3D) on the binary classification task (high ≥ 80% ee, low <80% ee). Cell entries are counts with per class accuracy percentages given.

The inclusion of SMOTE data produced more symmetric confusion matrices with similar balanced accuracy for both models. The additional synthetic variation helped the models to learn a decision boundary that discriminated better the minority class. However, in the case for the best-performing GBC(3D) model, this improvement came with a slight reduction in high class accuracy. Additionally, the biggest failure for both models still remained misclassifying low-*ee* as high.

The learning curves were fitted and plotted for the two models and are shown in **Figures 2.30** and **2.31**. With the addition of SMOTE the learning curves show similar performance to the initial curves shown in **Figures 2.24** and **2.25**, with a similar train-test gap. Thus, the models remain in a variance-limited regime, with substantial improvement requiring additional real data.

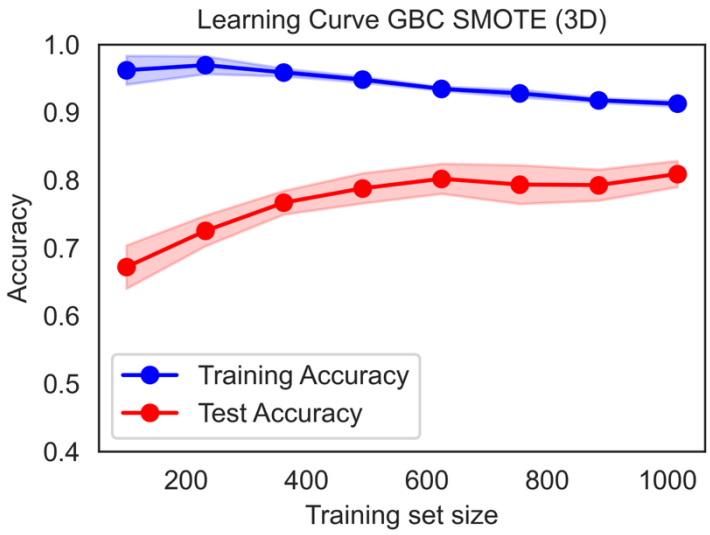


Figure 2.30: Learning curves for the GBC(3D) model with added SMOTE data. Blue markers show the CV training accuracy and the red markers show the CV test accuracy, reported as fractional accuracy (0-1 scale, where 1.0 corresponds to 100% accuracy).The shaded bands denote the SD.

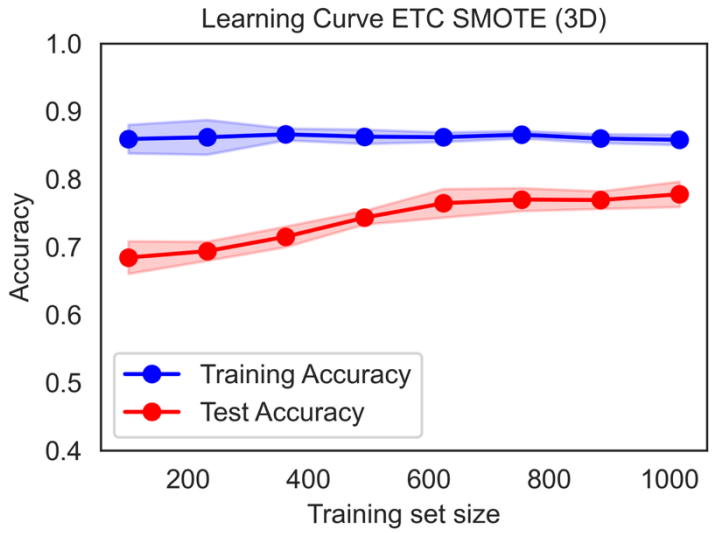


Figure 2.31: Learning curves for the ETC(3D) model with added SMOTE data. Blue markers show the CV training accuracy and the red markers show the CV test accuracy, reported as fractional accuracy (0-1 scale, where 1.0 corresponds to 100% accuracy). The shaded bands denote the SD.

The permutation feature importances of the two models are shown in **Figures 2.32** and **2.33**. For the GBC model, the feature importances became more concentrated with the difference in accuracy increasing from 0.01% (without SMOTE data) to 0.03% for the top feature with SMOTE data, indicating a smaller reliance on a smaller set of descriptors. For the ETC model the features remained at a 0.01% difference in

accuracy. Notably, the features for both models would now be reordered towards features that better distinguish the low *ee* class.

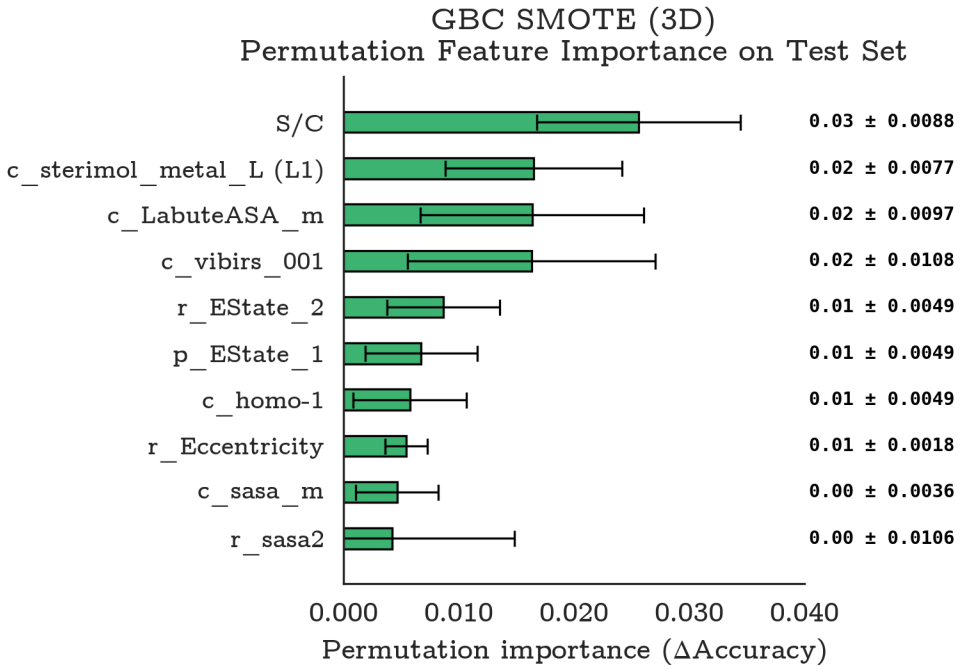


Figure 2.32: Permutation feature importance for the GBC model with additional SMOTE data trained on the 3D descriptor set. Bars show the mean increase in test accuracy when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

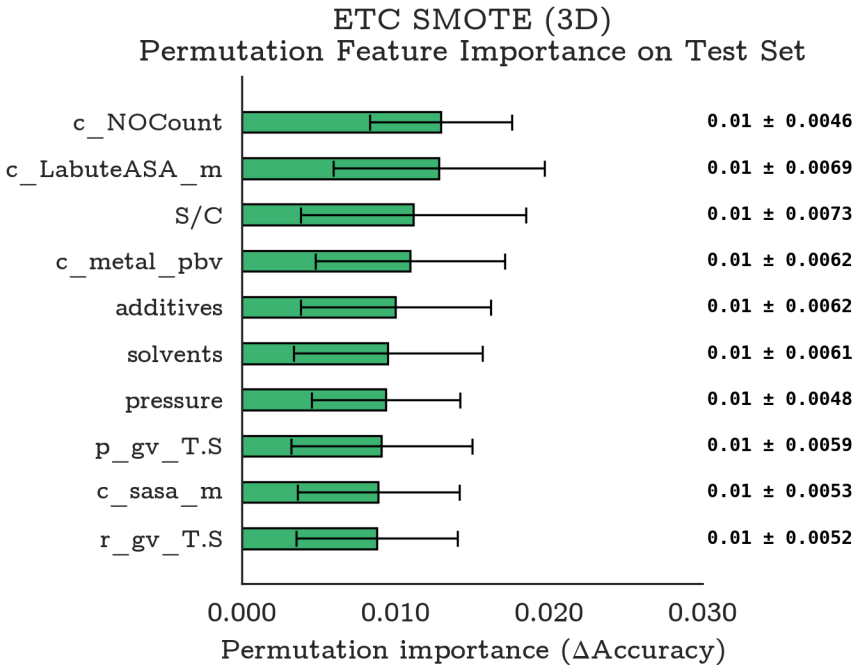


Figure 2.33: Permutation feature importance for the ETC model with additional SMOTE data trained on the 3D descriptor set. Bars show the mean increase in test accuracy when each feature's values are randomly permuted and error bars denote the SD across repeated permutations.

2.4 CONCLUSIONS

This chapter established a workflow for predicting *ee* in AH reaction of TSC=C from literature data. A curated dataset of 1,271 reactions was assembled. Three subsets of input features were constructed, 2D structural fingerprints (ECFPs), 3D physical-organic features from CREST/GFN2-xTB ground state geometries, and the combination of 2D+3D features, all including features for the experimental reaction conditions. Due to the reporting bias towards high-*ee* reactions present in the dataset, the prediction task was formulated as regression and classification tasks. For the regression task the models were trained to predict the continuous *ee* values and the MAE of predicted *ee* compared to true *ee* values were analysed. For classification task, the *ee* values were binned into a high *ee* class ($\geq$ 80% *ee*) and low *ee* class ($<$ 80% *ee*) and the accuracy of the models at predicting the correct class was analysed. Learning curves and permutation feature importance on the held-out test set were also reported. Additionally, class imbalance was addressed with SMOTE data.

The regression models found tree-based ensembles outperformed linear models with the ETR with combined 2D+3D features reaching a test MAE of 9.8 $\pm$ 1.2% *ee*, meeting a 10% *ee* MAE target. However, errors were heavily concentrated in the low-mid *ee* region due to the dataset bias, motivating the reformulation to a classification task.

In the classification models, across algorithms and descriptor sets, test accuracies were 76-81%, exceeding the majority class baseline of 67.9%. GBC with 3D descriptors achieved the top accuracy of 81.4 $\pm$ 1.1%. Confusion matrices revealed asymmetric prediction accuracy, high predictions were reliable (80-94% accuracy) but accuracy of low *ee* was limited (42-72%). 3D descriptors showed a predictive advantage, but the addition of 2D+3D features did not improve on 3D alone.

The inclusion of SMOTE data to the 3D classifier models GBC and ETC, increased the minority class accuracy. However, for the best-performing classifier GBC(3D), this came with a reduced high-class accuracy, emphasising the need for more real experimental data to increase model performance.

Across all models, the permutation feature importances showed no single dominant feature drove predictions. Instead *ee* was found to be governed by complementary steric, electronic and experimental reaction condition feature signals. The learning curves for the most predictive models indicated a variance limited region and predictive ability in both classification and regression tasks was improving on the largest available sample size. This further emphasises the need for more real experimental data for better predictive ability, specifically in the low-mid range *ee* region.

In conclusion, this chapter established a methodology that produced prediction for *ee* in both classification and regression tasks, with different options of input features, dependent on inclusion of QC calculations of ground-states. The regression models met the predefined 10% *ee* MAE benchmark, and the classification models indicate an accuracy of 81.4%. Both models show limitations in predictive ability around the low-mid range *ee* values and classes, where additional data is needed.

Given the current performance, the next step was to evaluate the models through their application to novel TSC=C AH reactions, validate the predictions experimentally, and assess suitability for integration into a real-time optimisation workflow.

3 Evaluating Machine Learning Models on Novel Asymmetric Hydrogenation Reactions of Tetra-substituted Olefins

3.1 INTRODUCTION

While ML has been applied in the literature to the prediction of experimental reaction outcomes, very few studies have validated these models on novel experimental reactions. Such evaluation is essential if ML models are to be applied in practice for the optimisation of unoptimised reactions, whether trained on literature data, where reactions are often reported under optimised conditions, or on HTE data, which may not reflect the same screening processes used in new studies.

In the previous chapter, ML models were developed using the available literature data on TSC=C reactions to predict both *ee* as a continuous value and the *ee* class, defined as high (≥80% *ee*) or low (<80% *ee*). These models demonstrated promising performance on held-out literature test sets, particularly the best-performing regression and classification models: ETR(2D+3D) and GBC(3D). However, their practical utility ultimately depends on how well they can predict outcomes for completely novel reactions.

To reflect how such models might be applied in practice, the test reactions in this chapter were designed to mimic an initial screening study, where conditions are not optimised, enantioselectivities are expected to be modest, and structurally diverse chiral ligands are assessed. For this purpose, 2,3-diphenyl-1H-inden-1-one (**S3.1**) was selected as the test substrate. Although the racemic hydrogenated product, 2,3-diphenylindan-1-one, has been prepared,[259] no method for its enantioselective synthesis has been reported. This made **S3.1** an ideal case study to evaluate the predictive capability of the models in the absence of prior training data.

Furthermore, evaluating the ML models on **S3.1** also provides broader insight into their potential to accelerate optimisation campaigns of relevance to drug discovery. Indanone derivatives are of particular interest due to their applications in pharmaceuticals, materials, and natural products,[260, 261] which has motivated prior studies on their racemic synthesis—most notably through palladium-catalysed transformations of vinyl benzoic acids.[259] One natural product of interest containing the diphenylindan-1-one motif is (+)-Isopaucifloral F (**Figure 3.1**).[262] Although racemic syntheses of (±)-Isopaucifloral F have been reported,[263] an enantioselective

synthesis has not yet been achieved. As such, if the generated models can accurately predict the *ee* of reactions involving **S3.1**, this would highlight their potential applicability in the development of intermediates to complex motifs such as (+)-Isopaucifloral F.

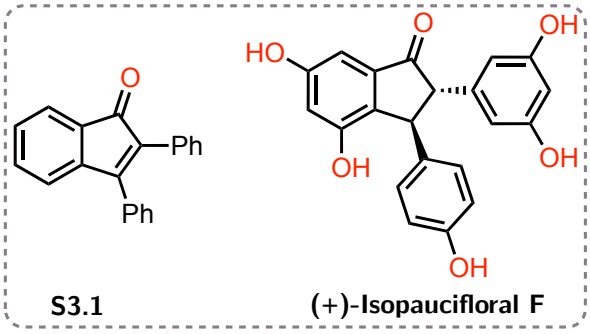


Figure 3.1: Structures of S3.1 and the natural product (+)-Isopaucifloral F.

In addition to testing the models on a completely novel substrate, it was also important to evaluate their performance on substrates already represented in the training dataset ('seen' substrates). This approach allows assessment of whether the models can reliably extend their predictions to new catalyst–substrate combinations, when the substrate itself has prior representation in the data. To this end, reactions of these substrates with chiral catalysts not previously reported for their transformations were carried out, providing a further test of the models' predictive capability

This chapter therefore presents the experimental evaluation of the ML models developed in this work, comparing predicted and observed outcomes for novel AH reactions, and assessing their potential and limitations for guiding future optimisation campaigns.

3.1.1 Chapter Aims

This chapter aimed to evaluate the predictive performance of the ML models developed in Chapter 2 on novel AH reactions of TSC=C. Specifically, the objectives were: (i) to design an experimental dataset comprising both 'unseen' and 'seen' substrates with a structurally diverse set of chiral ligands; (ii) to extract the relevant features from this dataset to ensure consistency with those used in model development; (iii) to generate predictions using the best-performing models, ETR(2D+3D) and GBC(3D); and (iv) to conduct the reactions experimentally and compare the predicted and experimental *ee* values in order to assess the models' reliability.

3.2 METHODS

The experimental dataset used to test the predictive ability of the ML models comprised the substrates and ligands shown in **Figure 3.2**. Substrate **S3.1** represents a novel 'unseen' substrate, whereas substrates **S3.2**–**S3.5** are 'seen' substrates for which reaction data already exist in the literature dataset. All ligands included in **Figure 3.2** are also present within the literature dataset.

To evaluate the predictive performance of the ML models on the 'unseen' substrate **S3.1**, conformational searching and geometry optimisation were performed using CREST at the GFN2-xTB level of theory, consistent with the procedure applied to the literature substrates. From the optimised structures, molecular fingerprints (2D) and physical organic features (3D) were extracted in alignment with the approach described in **Section 2.2.3**. The reaction conditions employed in the experimental reactions (outlined in detail in **Section 3.3**) were likewise encoded as input features in the same manner as in the original ML dataset.

The *ee* of the reactions was predicted using models ETR(2D+3D), GBR(2D+3D), GBR(3D), ETR(3D), ETC(3D), GBC(3D) and the SMOTE datasets ETC(3DSMOTE) and GBC(3DSMOTE) for comparison across feature sets and models.

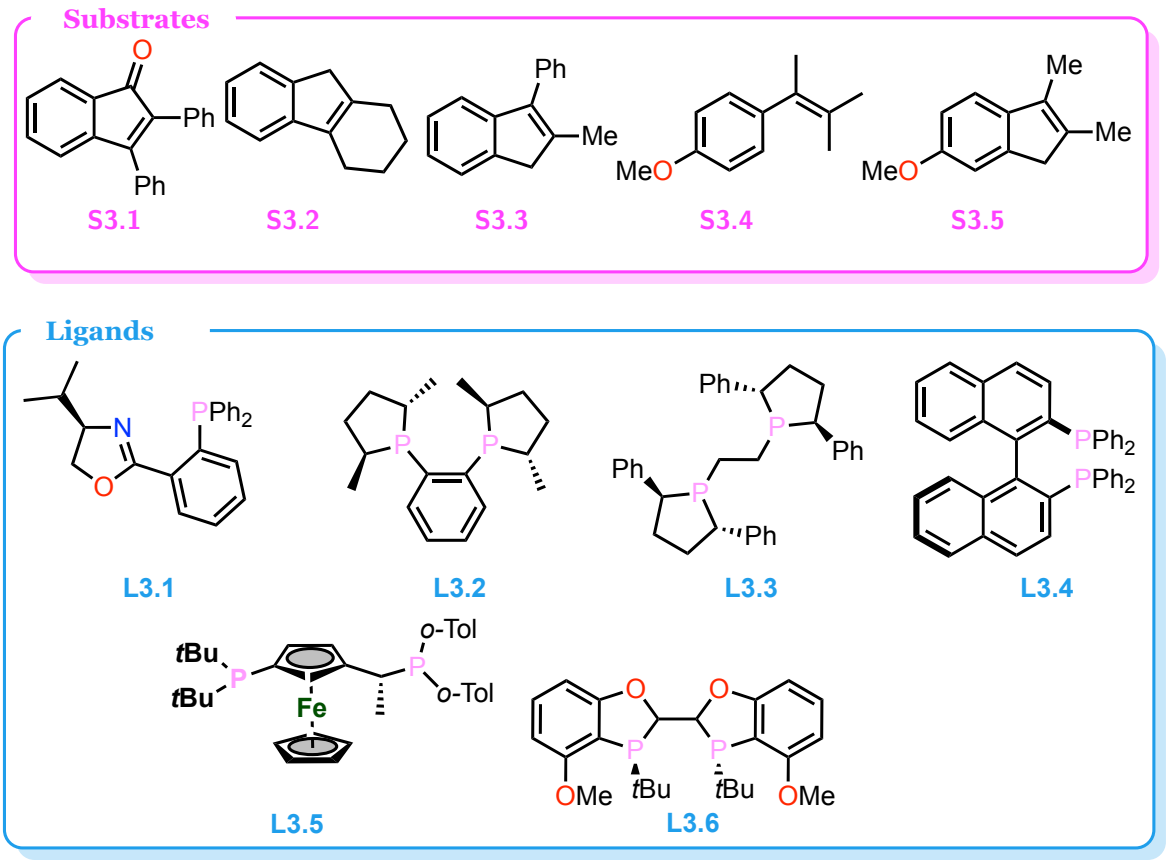


Figure 3.2: Substrates and ligands employed in the AH reactions of the experimental dataset.

3.3 RESULTS AND DISCUSSION

3.3.1 Prediction of Reactions of a Novel Substrate

To assess the predictive performance of the ML models on the 'unseen' substrate **S3.1**, reactions were carried out using four commercially available chiral ligands (**Figure 3.2**) that have previously been employed in AH reactions of TSC=C. It was hypothesised that catalysts prepared *in-situ* from these ligands would deliver adequate conversion, enabling their enantioselectivities to be evaluated. The frequency of occurrence of each ligand in the literature dataset is also shown in **Figure 3.3**. All four ligands have been used multiple times in prior studies and rank among the twelve most frequently employed ligands by reaction count (**Figure 2.4**, **Section 2.3**).

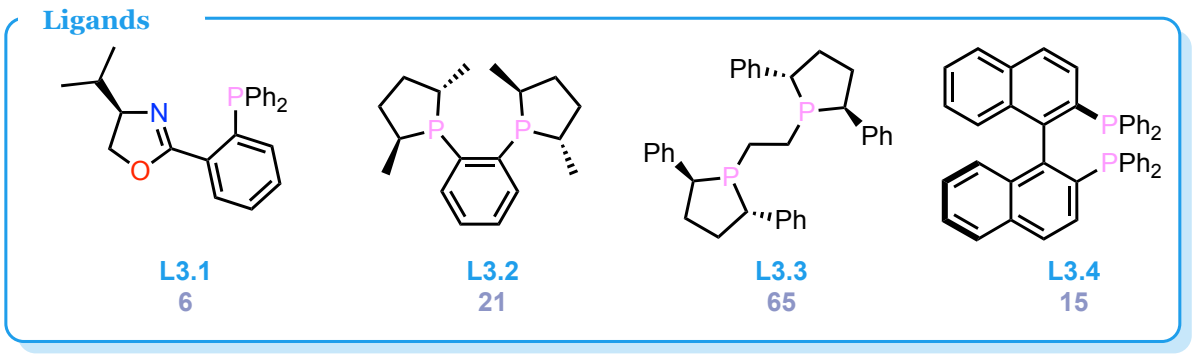
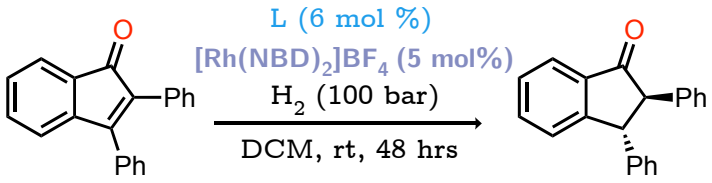


Figure 3.3. Ligands employed in the AH reactions of S3.1. The number of times each ligand appears in the literature dataset is reported below the ligand in purple.

The reactions were all conducted with 5 mol% $[Rh(NBD)_2]BF_4$, 6 mol% of a chiral ligand, in DCM at rt under 100 bar of H_2. The *ee* of the reactions were predicted using the best-performing regression model, ETR(2D+3D), while the *ee* class (high ≥ 80% *ee*, low < 80% *ee*) was predicted using the best-performing classification model, GBC(3D). The outcomes of these predictions (p.*ee* and p.class) and the newly obtained experimental *ee* values are presented in **Table 3.1**.

The reactions afforded moderate to high conversions, while the enantioselectivities were low to modest, ranging from 21% *ee* with **L3.1** to 73% *ee* with **L3.2**. The latter represents the first report of the enantioselective formation of this product. Further screening and optimisation may therefore yield a particularly useful catalytic system for this class of substrates.

Table 3.1: Results obtained for the AH reactions of S3.1 with ligands L3.1-L3.4, alongside the predicted *ee* values from ETR(2D+3D) and class prediction from GBC(3D).



R_N	Ligand	Conversion (%)	*d.r*	*ee*(%)	ETR(2D+3D) p.*ee*(%)	GBC(3D) p.class
3.1	**L3.1**	73	4:3	21	80	high
3.2	**L3.2**	95	100	73	75	high
3.3	**L3.3**	56	100	15	83	high
3.4	**L3.4**	34	100	31	58	high

Both the regression and classification models overestimated the enantioselectivity. The classification model, in particular, predicted all cases as belonging to the high class, despite the experimental *ee* values being below 80% and therefore belonging to the low class. The regression model gave one accurate prediction for reaction 3.2, estimating 75% *ee* compared with the experimental 73% *ee*, but performed poorly for the other reactions. This discrepancy likely reflects class imbalance in the training dataset, where 863 of the 1271 reactions exhibited an *ee* ≥ 80%. As a result, the models perform poorly when applied to low-*ee* reactions.

3.3.2 Predictions of Literature Substrates

Given the poor predictive performance of the models on low-*ee* reactions with an 'unseen' substrate, attention was redirected to predicting AH reactions on substrates already represented in the ML training dataset. The selected and synthesised substrates are shown in **Figure 3.4**, along with the number of reactions reported for each in the literature dataset. The ligands used for the reactions are shown in **Figure 3.5**: **L3.2** and **L3.4** were employed in the AH of **S3.1**, while **L3.5** and **L3.6** were also ligands in the literature dataset, used in 19 and 1 reactions respectively.

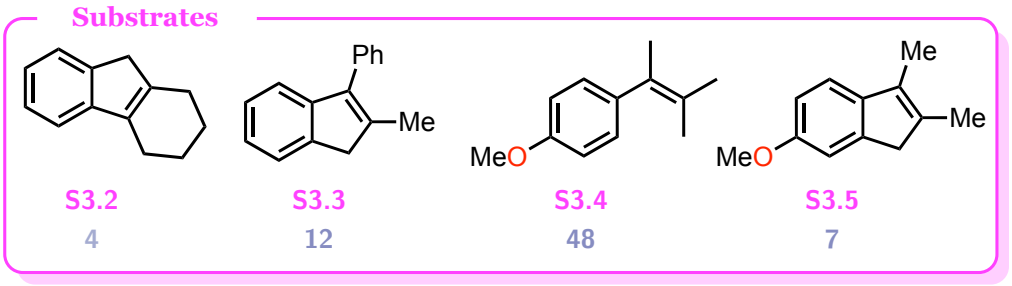


Figure 3.4: Substrates selected and synthesised for AH reactions. The number of reactions reported for each substrate in the literature dataset is indicated in purple.

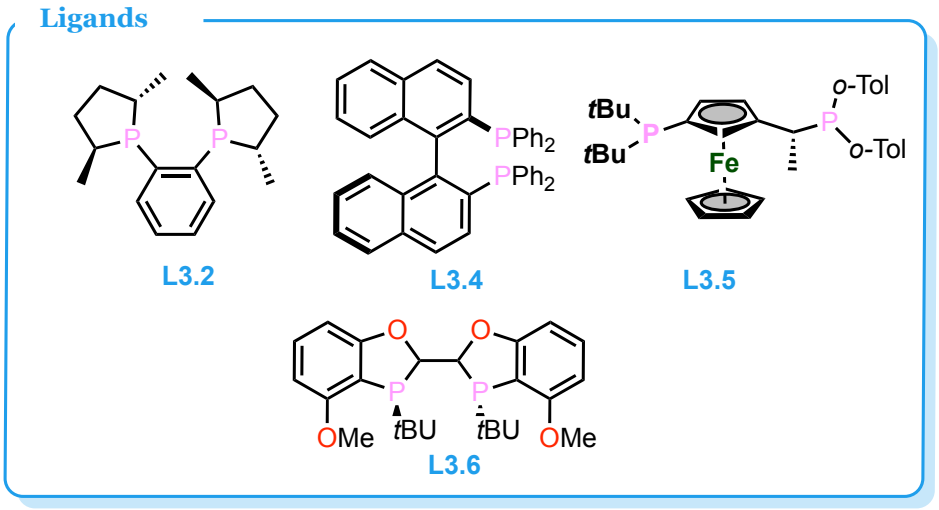


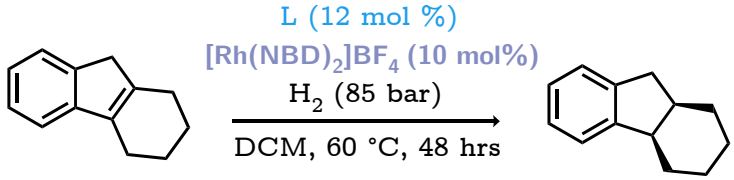
Figure 3.5: Ligands employed in the AH reactions of S3.2-3.5.

The use of substrates already represented in the ML training dataset was hypothesised to improve predictive performance, as the models could draw on prior reaction data. For example, **S3.4** is the second most frequently occurring substrate in the dataset when ranked by reaction count (**Figure 2.2**, **Section 2.3**). **S3.5** has previously been studied with ligands **L3.2** and **L3.6** in a report by Zhang (2017),[38] where it was screened against multiple ligands before anthryl-MeOBIPOP— a modified form of MeO-BIPOP (**L3.6**)—was identified as the optimal ligand. In that study, reactions were carried out using 10 mol% $[Rh(NBD)_2]BF_4$ and 12 mol% of the chiral ligand in DCM. These conditions therefore served as the basis for the reactions performed on the present substrates.

The *ee* of all reactions were again predicted using the best-performing regression model, ETR(2D+3D), while the *ee* class (high ≥ 80% *ee*, low < 80% *ee*) was predicted using the best-performing classification model, GBC(3D). The outcomes of these

predictions (p.*ee* and p.class) and the newly obtained experimental *ee* values for **S3.2** are presented in **Table 3.2**.

Table 3.2: Results obtained for the AH reactions of S3.2 with ligands L3.2, L3.4, L.5 and L3.5, alongside the predicted *ee* values from ETR(2D+3D) and class prediction from GBC(3D).



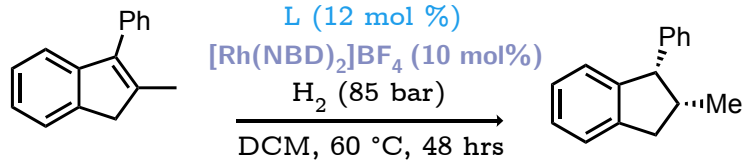
R_N	Ligand	Conversion (%)	*ee*(%)	ETR(2D+3D) p.*ee*(%)	GBC(3D) p.class
3.5	**L3.2**	>99	15	67	low
3.6	**L3.4**	>99	55	55	low
3.7	**L3.5**	0	N.D	75	low
3.8	**L3.6**	77	13	71	low

Although conversions were high for reactions 3.5, 3.6, and 3.8, the enantioselectivities were low, ranging from 13% to 55% *ee*, with the highest value of 55% obtained using **L3.4**. In contrast to the predictions for **S3.1**, the class predictions from GBC(3D) were fully accurate, correctly assigning all cases to the low class. The regression model, ETR(2D+3D), was less reliable as it substantially overestimated the *ee* for reactions 3.5 and 3.8, predicting 67% and 71% *ee* compared to the experimental 15% and 13% *ee*, respectively. However, it accurately predicted the outcome of reaction 3.6, which gave the highest *ee* for this substrate, at 55%.

The experimental results and predictions for **S3.3** (**Table 3.3**), again gave the highest *ee* with **L3.4** (78%), which also achieved full conversion for this substrate. Due to the low conversions obtained in reactions 3.11 and 3.12, *ee* values were only available for two reactions. Both were classified as high by GBC(3D), although their experimental *ee*s were low. The regression model, ETR(2D+3D), also failed to predict these values accurately, with absolute errors of 20 and 25%, respectively. Nevertheless, it is worth noting that although the predictions fell outside the 10% MAE achieved previously, the model did not default to reporting >90% *ee* (the *ee* values seen for the

majority of reactions in the literature dataset). This suggests the model is capable of assigning more moderate predictions, even if they are not yet sufficiently accurate.

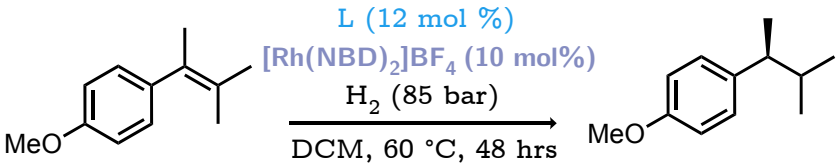
Table 3.3: Results obtained for the AH reactions of S3.3 with ligands L3.2, L3.4, L.5 and L3.5, alongside the predicted *ee* values from ETR(2D+3D) and class prediction from GBC(3D).



R_N	Ligand	Conversion (%)	*ee*(%)	ETR(2D+3D) p.*ee*(%)	GBC(3D) p.class
3.9	**L3.2**	39	45	65	**high**
3.10	**L3.4**	>99	78	53	**high**
3.11	**L3.5**	0	N.D	76	**high**
3.12	**L3.6**	19	N.D	72	**low**

A similar outcome was observed for **S3.4** (**Table 3.4**). The GBC(3D) model once again predicted all *ee* values to fall in the high class, despite the experimental results being low. The regression model performed best for the highest *ee* observed with this substrate, obtained using **L3.2** (69%), where the model predicted 59%. This falls within the 10% *ee* MAE benchmark defined for the regression models. However, for the lower *ee* values in reactions 3.14 and 3.16 (15% and 33%, respectively), the model overestimated the outcomes, predicting 60% and 68% *ee*.

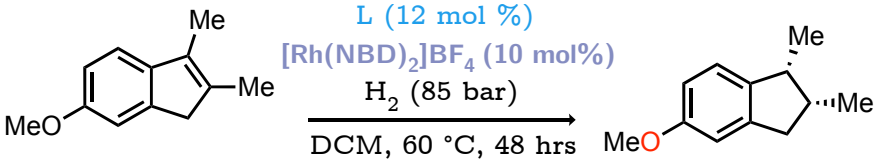
Table 3.4: Results obtained for the AH reactions of S3.4 with ligands L3.2, L3.4, L.5 and L3.5, alongside the predicted *ee* values from ETR(2D+3D) and class prediction from GBC(3D).



R_N	Ligand	Conversion (%)	*ee*(%)	ETR(2D+3D) p.*ee*(%)	GBC(3D) p.class
3.13	**L3.2**	>99	69	59	**high**
3.14	**L3.4**	>99	15	60	**high**
3.15	**L3.5**	0	N.D	72	**high**
3.16	**L3.6**	53	33	68	**high**

As noted previously, substrate **S3.5** has been studied with ligands **L3.2** and **L3.6** (Zhang, 2017). In the present work, experimental *ee* values were obtained for these pairings and, together with a novel reaction using **L3.5**, are summarised in **Table 3.5**. The results closely match the literature: **L3.2** and **L3.6** afforded 11% and 87% *ee* here, compared with 14% and 82% *ee* reported previously. Notably, the new reactions were conducted at higher hydrogen pressure (85 bar *vs* 60 bar) and shorter reaction time (48 h *vs* 96 h). Unfortunately, no conversion was observed for the novel reaction 3.18 with **L3.5**, precluding an evaluation of the ML models' performance on this substrate.

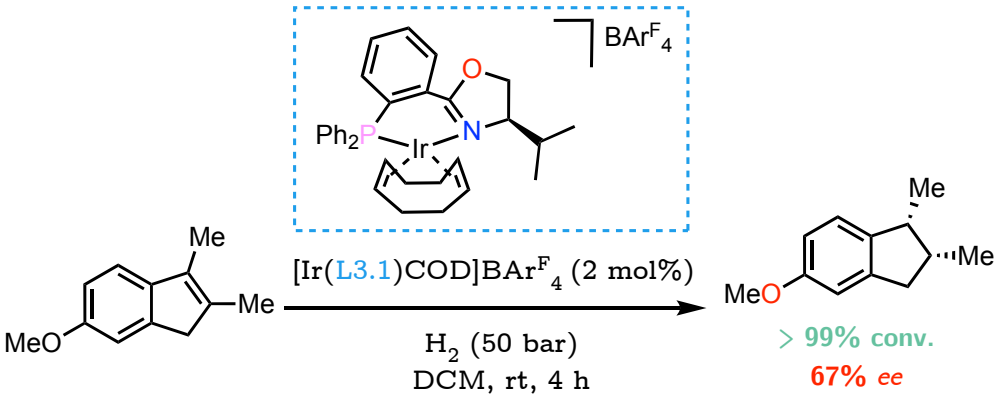
Table 3.5: Results obtained for the AH reactions of S3.5 with ligands L3.2, L.3.5 and L3.5, alongside the predicted *ee* values from the novel reaction 18 from ETR(2D+3D) and class prediction from GBC(3D).



R_N	Ligand	Conversion (%)	*ee*(%)	ETR(2D+3D) p.*ee*(%)	GBC(3D) p.class
3.17	**L3.2**	>99	11		
3.18	**L3.5**	0	N.D	74	high
3.19	**L3.6**	>99	87		

As this substrate could not be evaluated with the previously conducted reactions, attention turned to strategies that have shown success in the literature for unfunctionalised TSC=C substrates, particularly *P,N*-oxazoline ligands with [Ir(COD)][BAr^F_4] complexes. This approach had not yet been applied to the present substrate. Accordingly, the P,N-oxazoline ligand **L3.1** was employed with the complex [Ir(**L3.1**)(COD)]BAr^F_4, and the experimental results together with the model predictions are presented in **Table 3.6**. The reaction yielded >99% conversion and 67% *ee*. The *ee* class was again misclassified as high, whereas the true *ee* class was low (< 80% *ee*). However, the regression model ETR(2D+3D) predicted an *ee* of 73%, indicating a good performance of this model on this system.

Table 3.6: Results obtained for the AH of S3.6 with [Ir(L3.1)(COD)]BAr^F_4, alongside the predicted *ee* from ETR(2D+3D) and class prediction from GBC(3D).



R_N	Ligand	Conversion (%)	*ee*(%)	ETR(2D+3D) p.*ee*(%)	GBC(3D) p.class
3.20	**L3.1**	>99	**67**	**73**	**high**

This result aligns with the overall trend observed in the evaluation of the ML models on these substrates. When the experimental *ee* values are higher, the models predict them with reasonable accuracy. Across the full experimental dataset of 20 reactions, the MAE between predicted and true *ee* values was 31.3%, substantially higher than that reported from the literature test sets. This discrepancy reflects the fact that the training dataset, drawn primarily from optimised literature reactions, is not directly applicable in a real-time experimental setting. The accuracy of class predictions from GBC(3D) was also poor at 25%, as all experimental *ee*s fell into the low class; the only correct classifications were for **S3.2**. Excluding the first four reactions and recalculating the MAE for substrates already represented in the training dataset improved performance to 27.9% MAE *ee*. However, this remains considerably higher than the 10% *ee* MAE achieved by ETR(2D+3D) on the literature test set. This gap is likely due to the class imbalance in the training data, which contains a much higher proportion of high-*ee* reactions compared with the predominantly low-*ee* values in this experimental dataset.

For comparison of models and descriptor types the experimental dataset was also predicted by the GBR(2D+3D), ETR(3D), ETC(3D), GBC(3D) and the SMOTE

datasets ETC(3DSMOTE) and GBC(3DSMOTE) and the results are shown in **Table 3.7**.

Table 3.7: The prediction of *ee* values and class from models ETR(2D+3D), GBR(2D+3D), GBR(3D), ETR(3D), ETC(3D), GBC(3D) and the SMOTE datasets ETC(3DSMOTE) and GBC(3DSMOTE), alongside the true experimental *ee* values.

R_N	*ee*(%)	ETR (2D+3D) p.ee	GBR (2D+3D) p.ee	GBR (3D) p.ee	ETR (3D) p.ee	ETC (3D) p.ee	GBC (3D) p.ee	ETC (3DSMOTE) p.ee	GBC (3DSMOTE) p.ee
3.1	21	80	92	52	77	high	high	high	high
3.2	73	75	86	61	80	high	high	high	high
3.3	15	83	94	54	84	high	high	high	high
3.4	31	58	84	38	74	high	high	high	high
3.5	15	67	74	26	69	low	low	high	high
3.6	55	55	71	100	64	low	low	high	high
3.7	N.D	75	62	25	74	high	low	low	high
3.8	13	71	36	4	56	low	low	high	high
3.9	45	65	74	51	72	high	high	high	high
3.10	78	53	73	28	65	high	high	high	high
3.11	N.D	76	70	51	71	high	high	high	high
3.12	N.D	72	52	24	57	low	low	high	high
3.13	69	59	84	74	74	high	high	high	high
3.14	15	60	83	54	66	high	high	high	high
3.15	N.D	72	80	83	77	high	high	high	high
3.16	33	68	49	45	60	high	high	high	high
3.17	11								
3.18	N.D	74	65	79	76	high	high	high	high
3.19	87								
3.20	67	73	49	43	57	low	high	high	high

Overall, the models tended to overestimate enantioselectivity and underestimate the number of poor outcomes. Most regression models consistently predicted relatively high *ee* values, typically in the 60–90% range, even when the true experimental *ee* was low. For ETC(3D SMOTE) and GBC(3D SMOTE), nearly all reactions were classified as high regardless of the experimental value, suggesting that the use of SMOTE worsened class calibration in this case. GBR(3D) was the only model that occasionally identified low-*ee* reactions, consistent with its status as the best-performing model in the literature dataset. Notably, while ETR(2D+3D) performed best on the literature

dataset, in the experimental dataset GBR(3D) achieved the lowest MAE of 24% *ee* for this experimental dataset.

GBR(3D) achieved an MAE of 14.4 ± 0.6% *ee* on the test sets of the literature dataset. Although the MAE obtained for the experimental dataset was much higher, examination of the predictions reveals that this model outperformed the others on low *ee* reactions. In particular, reactions 3.4, 3.5, 3.8, 3.9, and 3.16 (all of which gave low experimental *ee*) were predicted with reasonable accuracy by GBR(3D). This suggests that, despite its lower overall MAE on the literature dataset, GBR(3D) may provide a more reliable alternative when applied to experimental reactions characterised by modest enantioselectivity.

3.4 CONCLUSIONS AND FUTURE WORK

This chapter evaluated the predictive performance of the ML models developed in Chapter 2 on novel AH reactions of TSC=C. The experimental dataset comprised both a completely 'unseen' substrate (**S3.1**) and 'seen' substrates (**S3.2-3.6**), tested with a diverse set of chiral ligands.

Across the dataset, the models overestimated enantioselectivity, particularly for reactions with low experimental *ee*. The best-performing regression model on the literature dataset ETR(2D+3D) tended to predict values in the 60-90% *ee* range, even when the true *ee* values were much lower, while the classification models frequently misclassified low-*ee* reactions as belonging to the high *ee* class. The outcomes highlight the influence of class imbalance in the training dataset, which was dominated by optimised literature reactions with *ee* ≥80%.

Despite this limitation, certain encouraging trends were observed. The models performed more reliably when experimental *ee*s were relatively high, with some predictions falling in the 10% *ee* MAE benchmark. For example, ETR(2D+3D) accurately predicted the *ee* of reaction 3.2 (73% vs 75%), reaction 3.6 (55%), and reaction 3.20 (67% vs 73%). The GBR(3D) model showed better generalisation to low-*ee* reactions, achieving the lowest overall MAE of 24% on the experimental dataset.

Future improvements would require further balancing the training dataset with a broader distribution of low *ee* examples, incorporating negative 'failed' reactions to counteract the reporting bias that exists in the dataset, and expanding the diversity of substrates and ligands included. Such improvements could enable the development of models with improved calibration across the full range of experimental outcomes and enhance their practical use for guiding optimisation campaigns of this challenging transformation.

3.5 EXPERIMENTAL

3.5.1 General Considerations

Unless otherwise stated all manipulations were carried out under an inert atmosphere (N_2 or Ar) using standard Schlenk techniques or employment of an MBraun glovebox. Glassware was dried in an oven at 180 °C overnight before use.

Unless otherwise stated, all reagents were purchased from Merck or Fluorochem and used as purchased. Laboratory grade toluene and tetrahydrofuran (THF) were dried over sodium/benzophenone and distilled prior to use. MeOH, DCM and ethyl acetate (EtOAc) were refluxed over calcium hydride (CaH) and distilled onto activated 3Å molecular sieves. Room temperature (rt) refers to 298 K. Ethylene diamine was also dried over activated 3Å molecular sieves.

High pressure hydrogen reactions were conducted in a HEL cat24 high pressure reactor vessel.

^{1}H, $^{13}C\{^{1}H\}$ and $^{31}P\{^{1}H\}$ NMR spectra were recorded on a Brucker 400 MHz or 500 MHz spectrometers in $CDCl_3$ and ^{1}H and $^{13}C\{^{1}H\}$ NMR chemical shifts are reported relative to $CDCl_3$ at δ_H 7.26 ppm and δ_C 77.16 ppm. Coupling constants (J) are reported in Hertz (Hz). Multiplicities are indicated by: br s (broad singlet), s (singlet), d (doublet), t (triplet), q (quartet), p (pentet) and m (multiplet).

Chiral supercritical fluid chromatography (SFC) analysis was performed on a Waters ACQUITY UPC2 System with DAICEL CHIRALPAK IC, IJ, IK or IG columns (4.6 × 250 mm, 3,0 μm) in a mixed solvent system of supercritical carbon dioxide (CO_2) and MeOH. A system backpressure of 138 bar was used in all cases.

3.5.2 General Procedure for Asymmetric Hydrogenation Reactions

In a nitrogen-filled glovebox a 4 mL vial equipped with a magnetic stirrer bar was charged with either $[Rh(NBD)_2]BF_4$ (5 mol%, 0.005 mmol, 1.86 mg), chiral ligand (6 mol%, 0.006 mmol); $[Rh(NBD)_2]BF_4$ (10 mol%, 0.010 mmol, 3.74 mg), chiral ligand (12 mol%, 0.012 mmol) or 2 mol% [Ir(L3.1)COD]$BAr^F{}_4$ and DCM (0.5 mL). The resultant mixture was stirred for 30 min before the substrate (0.1 mmol, 17 mg) was added. The charged vial was placed into a stainless-steel autoclave (HEL cat24 high

pressure reactor vessel). The autoclave was then purged with H_2 three times and charged to the indicated pressure and temperature. The mixture was stirred for 24 h before it was cooled to room temperature and depressurised. The resultant mixture was directly passed through a short plug of silica with EtOAc/petreoleum ether to give the desired product, evidenced by ^{1}H NMR accordance with the racemate product to give the conversion of starting material to product. The *ee* of the product was determined by chiral SFC analysis.

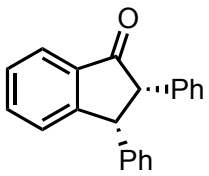
3.5.3 General Procedure for Racemic Hydrogenation Reactions

To 4 mL vial the relevant substrate (0.1 mmol), palladium on carbon (Pd/C) 10 wt% (0.01 mmol, 2 mg) and 500 uL EtOAc was charged. The vial was places in high pressure autoclave (HEL cat24 high pressure reactor vessel). The autoclave was pressurised to 100 bar H_2 and the reaction was left stirring for 18 hours. After carefully depressurising the autoclave the reaction was passed through celite, the volatiles were subsequently removed at reduced pressure. Purification by column chromatography (petroleum ether/ethyl acetate, 4:1) obtained the hydrogenated product.

3.5.4 Characterisation of Hydrogenated Product 2,3-Diphenyl-indan-1-one

2,3-diphenyl-indan-1-one (P3.1)

(S,R)/(R,S)

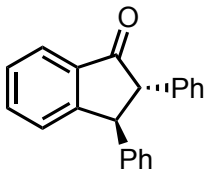


^{1}H NMR (400 MHz, $CDCl_3$): δ 7.97 (d, J = 7.5 Hz, 1H), 7.7 – 7.66 (m, 1H), 7.54 – 7.49 (m, 1H), 7.44 - 7.41 (m, 1H), 7.36 – 7.24 (m, 2H), 7.05 – 6.95 (m, 4H), 6.81 – 6.75 (sm 1H), 6.65 – 6.57 (m, 1H), 5.07 (d, J = 8.2 Hz, 1H), 4.49 – 4.36 (m, 1H).

^{13}C{^{1}H} NMR (101 MHz, $CDCl_3$): δ 206.4, 155.8, 140.7, 138.0, 137.1, 135.8, 130.5, 129.8, 128.75, 128.27, 127.85, 126.96, 124.2, 61.2, 52.0.

Chiral SFC Analysis: CHIRALPAK IC (CO_2:MeOH, 90:10, 1.25 mL min^{-1}, 40 °C, 240 nm) tR = 8.8, 10.6 minutes.

(S,S)/(R,R)



^{1}H NMR (400 MHz, $CDCl_3$): δ 7.89 (d, J = 7.7 Hz, 1H), 7.66 – 7.62 (m, 1H), 7.51 – 7.47 (m, 1H), 7.34 – 7.29 (m, 7H), 7.12 – 7.08 (m, 2H), 4.58 (d, J = 4.8 Hz, 1H), 3.82 (d, J = 4.8 Hz, 1H).

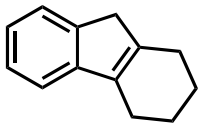
^{13}C{^{1}H} NMR (101 MHz, $CDCl_3$): δ 138.3, 136.4, 135.3, 128.8, 128.7, 128.2, 128.1, 127.8, 127.0, 126.5, 123.9, 64.4, 54.7.

Chiral SFC Analysis: CHIRALPAK IC (CO_2:MeOH, 90:10, 1.25 mL min^{-1}, 40 °C, 240 nm) tR = 6.1, 6.8 minutes.

MS ES+: 284.10 ([$C_{21}H_{16}O$]

3.5.5 Preparation of Substrates

1,2,3,4-tetrahydro-9H-fluorene (S3.2)



Synthesised according to reported procedure.[264]

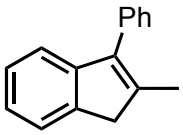
In a 250 mL flame dried Schlenk flask Fluorene (3.33 g, 19.33 mmol) was dissolved in 20mL of THF and 20 mL ethylene diamine under an argon atmosphere. The solution was cooled to 0 °C, and lithium (2 g, 290 mmol) was added in 500 mg portions every 10 min. The mixture was brought to rt and left to stir for 18 h. The reaction was carefully quenched with water and extracted with cyclohexane. Organic layers were dried with magnesium sulfate ($MgSO_4$) and the solvent was removed at reduced pressure. Purification by column chromatography (petroleum ether) obtained 1,2,3,4-tetrahydro-9H-flourene as a white crystalline solid. Yield 2.45 g (70%).

Spectroscopic analysis aligns with literature.[264]

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.43 (d, J = 7.3 Hz, 1H, CH=), 7.33 – 7.29 (m, 1H, CH=), 7.26 – 7.24 (m, 1H, CH=), 7.19 – 7.16 (m, 1H, CH=), 3.28 (s, 2H, CH_2), 2.54 – 2.43 (m, 4H, CH_2), 1.89 - 1.82 (m, 4H, CH_2) ppm.
^{13}C{^{1}H} NMR (101 MHz, $CDCl_3$): δ 146.3, 142.8, 141.5, 136.1, 125.9, 123.6, 123.3, 117.5, 40.6, 26.0, 23.3, 22.7, 22.3 ppm.

HRMS ES+: 171.1167 ($[M+H]^+$, $[C_{13}H_{14}]$)

2-Methyl-3-phenyl-1H-indene (S3.3):



Synthesised according to reported procedure.[265]

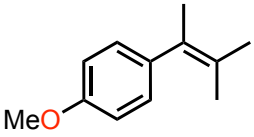
A Schlenk flask was charged with Mg (4.3g, 180 mmol) and was flame dried under vacuum. 250 mL THF and a catalytic amount of iodine were added and the mixture was stirred for 2 h. Bromobenzene (3.79 mL, 86 mmol) was then added to the mixture at 0 °C. The reaction was warmed to rt and then stirred at reflux for 18 h. The reaction mixture was filtered away from the excess Mg. 2-methyl-1-indanone (2.7 mL, 18.46 mmol) was added to the prepared Grignard dropwise at 0 °C. The reaction was then warmed to rt and stirred at reflux for 18 h. The reaction was cooled to room temperature and quenched with water (40 mL). HCl (0.1M, 40 mL) and Et_2O (40 mL) were added. The organic layers were extracted with Et_2O (3 × 20 mL), washed with saturated sodium hydrogen carbonate ($NaHCO_3$), brine, water and dried over $MgSO_4$. Purification by column chromatography (hexane: EtOAc ; 95:5) obtained 2-methyl-3-phenyl-1H-indene as a white solid (2.1 g, 55% yield).

Spectroscopic analysis aligns with literature.[265]

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.54 – 7.44 (m, 5H, CH=), 7.42 – 7.39 (m, 1H, CH=), 7.38 (ddd, J = 7.6, 3.7, 1.5 Hz, 1H, CH=), 7.20 (ddd, J = 7.3, 5.9, 2.6 Hz, 1H, CH=), 3.50 (s, 2H, CH_2), 2.19 (s, 3H, CH_3) ppm.
^{13}C{^{1}H} NMR (126 MHz, $CDCl_3$): δ 146.7, 142.8, 141.0, 139.0, 135.9, 129.5, 128.8, 127.4, 126.5, 124.4, 123.8, 119.4, 43.5, 15.2 ppm.

MS ES+: 206.1 ($[C_{14}H_{16}]$)

1-Methoxy-4-(3-methylbut-2-en-2-yl)benzene (S3.4):



Synthesised according to reported procedure using Wittig olefination.[266] Isopropyltriphenylphosphonium iodide (4.0 g, 9.26 mmol) potassium bis(trimethylsilyl)amide (KHMDS) (1.8 g, 9.26 mmol) and toluene (40 mL) were charged to a reaction vessel. The reaction was stirred at rt for 10 minutes until the mixture turned red. A solution of 1(4-methoxyphenyl)ethan-1-one (772 mg, 5.14 mmol) in toluene (40 mL) was added at rt. The reaction was heated to reflux for 2 h and then stirred at rt overnight. Volatiles were then removed at reduced pressure, diluted with Et_2O and filtered through silica. Purification by column chromatography (hexane: EtOAc ; 7:3) yielded 1-methoxy-4-(3-methylbut-2-en-2-yl)benzene (361 mg, 80%) as a coloourless oil.

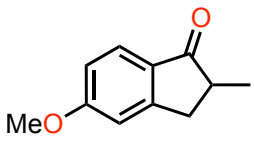
Spectroscopic analysis aligns with the literature.[266]

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.06 (d, J = 8.8 Hz, 2H, CH=), 6.86 (d, J = 8.8 Hz, 2H, CH=), 3.81 (s, 3H, CH_3-O), 2.16 – 1.89 (m, 3H, CH_3), 1.88 – 1.74 (m, 3H, CH_3), 1.68 – 1.57 (m, 3H, CH_3) ppm.

^{13}C{^{1}H} NMR (101 MHz, $CDCl_3$): δ 157.7, 137.9, 129.5, 127.1, 113.4, 55.3, 22.2, 21.0, 20.7 ppm.

MS ES+: 176.10 ([$C_{12}H_{16}O$])

5-methoxy-2-methyl-2,3-dihydro-1*H*-indenone:



Synthesised according to reported procedure.[202]

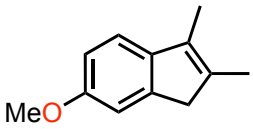
A Schlenk flask was charged with 5-methoxy-1-indanone (1.25g, 7.69 mmol) and THF (70 mL). The solution was cooled to − 40 °C and a solution of LiHMDS (1.7g, 10 mmol, 10 mL) in THF (40 mL) was added dropwise. The reaction mixture was stirred

at − 40 °C for 2 h, after which methyl iodide (MeI) (0.67 mL) was added, and the reaction mixture was stirred at 0 °C overnight. The reaction was diluted with aqueous ammonium chloride (NH_4Cl) and the volatiles were removed at reduced pressure. The residue was extracted with toluene. The organic phases were combined, washed with water (3 × 20 mL) and brine (3 × 20 mL), dried over $MgSO_4$ and volatiles were removed at reduced pressure. Purification by column chromatography (hexane: EtOAc ; 7:3) yielded 5-methoxy-2-methyl-2,3-dihydro-1*H*-indenone as a white solid (790 mg, 59% yield).

Spectroscopic analysis aligns with literature.[202]

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.75 – 7.55 (m, 1H, CH=), 7.01 – 6.75 (m, 2H, CH=), 3.88 (s, 3H, CH_3-O), 3.47 – 3.24 (m, 1H, CH), 2.85 – 2.58 (m, 2H, CH_2), 1.30 (d, J = 7.3 Hz, 3H, CH_3) ppm.

6-Methoxy-2,3-dimethyl-1*H*-indene (S3.5):



Synthesised according to reported procedure.[202]

To a solution of the prepared 5-methoxy-2-methyl-2,3-dihydro-1*H*-indenone (790 mg, 4.48 mmol) in THF (20 mL) at 0 °C was added methyl lithium (MeLi) (4.2 mL, 6.72 mmol) was added. The solution was then stirred at rt for 18 h. The reaction was quenched with saturated NH_4Cl dropwise. Subsequently, Et_2O and water were added. The organic phase was extracted with Et_2O. The combined organic phases were washed with sodium chloride (NaCl), dried over $MgSO_4$ and volatiles were removed at reduced pressure. The resulting oil was extracted with toluene (20 mL). *p*-Toluenesulfonic acid (5 mg) was added and the mixture was heated to reflux overnight. The solution was allowed to cool to rt, Et_2O and saturated $NaHCO_3$ was added. The organic phase was separated, washed with brine, dried over $MgSO_4$ and volatiles were removed at reduced pressure. Purification by column chromatography (petroleum ether: EtOAc ; 9:1) yielded 6-Methoxy-2,3-dimethyl-1*H*-indene as a yellow oil (470 mg, 60% yield).

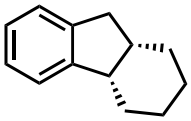
Spectroscopic analysis aligns with literature.[202]

^{1}H NMR (400 MHz, $CDCl_3$): δ 2.01 (s, 3H, CH_3), 2.03 (s, 3H, CH_3), 3.23 (s, 2H, CH_2), 3.82 (s, 3H, CH_3-O), 6.82 (dd, 1H, CH=, J= 8.2, J= 2.4 Hz), 6.98 (d, 1H, CH=, J= 2.4 Hz), 7.09 (d, 1H, CH=, J= 8.2 Hz) ppm.

MS ES+: 174.10 ([$C_{12}H_{14}O$])

3.5.6 Characterisation of Hydrogenated Products

cis-1,2,3,4,4a,9,9a-heptahydro-4aH-flourene (P3.2):



Spectroscopic analysis aligns with literature.[34]

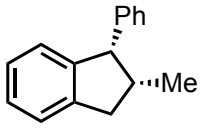
^{1}H NMR (400 MHz, CDCl3): δ 7.30 – 7.24 (m, 1H), 7.19 (dtt, J = 15.5, 6.4, 3.1 Hz, 3H), 3.14 (q, J = 5.8 Hz, 1H), 2.90 (dd, J = 15.1, 6.6 Hz, 1H), 2.63 (dd, J = 15.1, 4.4 Hz, 1H), 2.48 (dqd, J = 8.4, 6.3, 4.6 Hz, 1H), 1.97 – 1.86 (m, 1H), 1.81 (dddd, J = 13.9, 9.8, 5.6, 3.8 Hz, 1H), 1.69 – 1.55 (m, 1H), 1.56 – 1.19 (m, 5H).

^{13}C{^{1}H} NMR (101 MHz, CDCl3): δ 147.1(CH_{Ar}), 144.0 (CH_{Ar}), 126.1 (CH_{Ar}), 125.9 (CH_{Ar}), 125.3 (CH_{Ar}), 123.0 (CH_{Ar}), 44.0 ($C_{Ar}CH$), 39.8 ($C_{Ar}CHCH$), 37.4 ($C_{Ar}CH_2$), 28.1 (CH_2), 27.3 (CH_2), 24.0 (CH_2), 22.7 (CH_2) ppm.

HRMS: m/z 145.0755 $(M+2(CH_3COO))^{-2}$, 217.122 $(M+HCOO)^-$, 231.1444 $(M+CH3COO)^-$, calcd for $C_{13}H_{16}$ 172.1247.

Chiral SFC Analysis: CHIRALPAK IJ (CO_2:MeOH, 99.5:0.5, 2.50 mL min^{-1}, 40 °C) tR = 3.7 (major), 4.0 (minor) minutes.

cis-1-phenyl-2-methylindane (P3.3):



Spectroscopic analysis aligns with literature.[202]

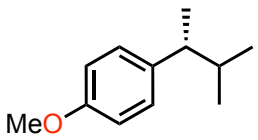
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.33 – 7.27 (m, 3H), 7.25 – 7.11 (m, 5H), 7.03 – 6.97 (m, 1H), 4.39 (d, J = 7.9 Hz, 1H, CH-Ph), 3.07 (dd, J = 15.4, 7.4 Hz, 1H, CH_2), 2.93 – 2.79 (m, 1H, CH-Me), 2.69 (dd, J = 15.4, 7.4 Hz, 1H, CH_2), 0.72 (d, J = 7.0 Hz, 3H, CH_3) ppm.

^{13}C{^{1}H} NMR (126 MHz, $CDCl_3$): δ 146.4, 144.2, 141.7, 129.6, 128.1, 126.8, 126.6, 126.3, 125.6, 124.7, 55.4, 39.9, 39.6, 16.9 ppm.

Chiral SFC Analysis: CHIRALPAK IG (CO_2:MeOH, 95:5, 2.50 mL min^{-1}, 40 °C), tR = 2.1 (major), 2.6 (minor) minutes.

MS ES+: 208.30 ([$C_{16}H_{16}$])

1-Methoxy-4-(3-methylbutan-2-yl)benzene (P3.4):



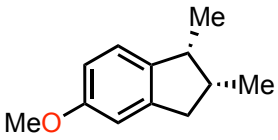
Spectroscopic analysis aligns with literature.[202]

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.07 (d, J = 8.6 Hz, 2H, CH=), 6.83 (d, J = 8.7 Hz, 2H, CH=), 3.79 (s, 2H, CH_3), 2.47 – 2.29 (m, 1H, CH), 1.80 – 1.63 (m, 1H, CH), 1.21 (d, J = 7.1 Hz, 3H, CH_3), 0.91 (d, J = 6.7 Hz, 3H, CH_3), 0.75 (d, J = 6.7 Hz, 3H, CH_3) ppm.

^{13}C{^{1}H} NMR (101 MHz, $CDCl_3$): δ 128.6, 113.5, 55.4, 46.1, 34.7, 21.2, 20.3, 19.0 ppm.

Chiral SFC Analysis: CHIRALPAK IG (CO_2:MeOH, 99.5:0.5, 2.50 mL min^{-1}, 40 °C, 219 nm), tR = 6.2, 7.1 minutes.

5-Methoxy-1,2-dimethyl-2,3-dihydro-1H-indene (P3.5):



Spectroscopic analysis aligns with literature.[202]

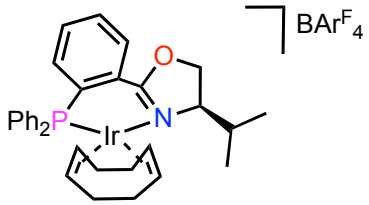
^{1}H NMR (500 MHz, CDCl3): δ 7.07 (d, J = 8.1 Hz, 1H, CH=), 6.79 – 6.68 (m, 3H, CH=), 3.79 (d, J = 1.0 Hz, 3H, CH_3), 3.11 (p, J = 7.1 Hz, 1H, CH), 2.95 (dd, J = 14.6, 6.5 Hz, 1H, CH), 2.62 – 2.49 (m, 2H, CH_2), 1.11 (dd, J = 7.1, 1.0 Hz, 3H, CH_3), 0.98 (dd, J = 6.9, 1.1 Hz, 3H, CH_3) ppm.

MS ES+: 176.10 ([$C_{12}H_{16}O$])

Chiral SFC Analysis: CHIRALPAK IG (CO_2:MeOH, 99.5:0.5, 2.50 mL min^{-1}, 40 °C, 279 nm), tR = 8.7, 9.4 minutes.

3.5.7 Synthesis of Iridium Complex

2-(2-{Diphenylphosphanylphenyl)-4-isopropyl-4,5-dihydrooxazole-(η^4-1,5-cyclooctadienyl)iridium(I)Tetrakis(3,5-bis(trifluoromethyl)phenyl)borate ([Ir(L3.1)(COD)]BAr^F_4)



In A Schlenk flask under an argon atmosphere, a solution of ligand (219 mg, 0.59 mmol) and $[Ir(COD)Cl]_2$ (198 mg, 0.295 mmol) in DCM (10 mL) was stirred at 40 °C for 1.5 h. After cooling to rt, a suspension of $NaBAr^F_4$ (1.08 g, 1.22 mmol) in degassed H_2O (40 mL) was added, and the resulting mixture was stirred at rt for 4 h. The organic layer was extracted with DCM and dried over $MgSO_4$. Purification by column chromatography (Petroleum ether:DCM ; 8:2) yielded the complex as a bright red powder.

Spectroscopic analysis aligns with literature.[267]

^{1}H NMR (500 MHz, $CDCl_3$): δ 8.18 – 8.06 (m, 1H, CH_{Ar}), 7.76 – 7.68 (m, 8H, CH_{Ar}), 7.65 – 7.54 (m, 5H, CH_{Ar}), 7.53 – 7.44 (m, 6H, CH_{Ar}), 7.47 – 7.38 (m, 3H, CH_{Ar}), 7.15 – 6.96 (m, 2H, CH_{Ar}), 5.06 – 4.95 (m, 1H, CH of COD), 4.91 (p, J = 7.2 Hz, 1H, CH of COD), 4.42 (dd, J = 9.6, 3.0 Hz, 1H, CH_2) 4.33 (t, J = 9.5 Hz, 1H, CH), 4.11 (dt, J = 9.5, 3.1 Hz, 1H, CH), 3.35-3.30 (m, 1H, CH of COD), 3.10 – 3.03 (m, 1H, CH of COD), 2.63 -2.38 (m, 4H, CH_2 of COD), 2.14- 2.04 (m, 1H, CH_2 of COD), 2.04-1.93 (m, 2H, CH_2 of COD), 1.78-1.65 (m, 1H, CH_2 of COD), 1.50-1.44 (m, 1H, $CH(CH_3)_2$), 0.83 (d, J = 7.1 Hz, 3H, CH_3), -0.10 (d, J = 6.7 Hz, 3H, CH_3).

^{31}P{^{1}H} NMR (203 MHZ, $CDCl_3$): δ 15.88

^{13}C NMR (126 MHz, $CDCl_3$): δ 162.1, 134.94, 134.33, 134.20, 134.14, 134.08, 134.02, 133.32, 133.24, 132.94, 132.57, 132.32, 130.05, 129.96, 129.11, 129.02, 125.8, 123.6, 117.6, 97.2, 93.6, 70.7, 68.4, 64.2, 63.8, 36.3, 33.1, 32.3, 28.7, 26.6, 18.7, 12.4.

4 Applying Bayesian Optimisation to Asymmetric Hydrogenation Reactions of Tetra-substituted Olefins

4.1 INTRODUCTION

In the previous chapters, ML models were developed and experimentally evaluated on their ability to predict the *ee* values of AH reactions of TSC=C. Although, the resultant MAEs from the literature-trained ML models indicated strong performance, the models failed to predict the *ee* sufficiently on novel substrates in real experimental settings, specifically for low-*ee* values, thereby limiting their practical use in reducing costly trial-and-error experimentation.

Given the vastness of the chemical space associated with this reaction class, it is likely thousands, or even tens of thousands of additional data points would be required to yield a generalisable and accurate model for the prediction of *ee* values. For this reason, an alternative optimisation strategy that is both efficient and applicable is needed.

The method chosen for this was Bayesian optimisation (BO) which was applied to a single substrate, with the objective of identifying the optimal reaction conditions that maximise both *ee* and yield.

4.1.1 Acknowledgment of Contributing Researchers

The coding of the BO framework used in this chapter was conducted with the assistance of Toby Lewis-Atwell.

4.1.2 Bayesian Optimisation

BO is an iterative design strategy used for the global optimisation of black-box functions. It has proven highly effective in tuning ML models and has found increasing applications in the field of chemistry.[268-274] Recently, BO has been used for optimising chemical reactions and has shown excellent performance.[273] In this context, it has been applied to iteratively suggest new reaction conditions and aid the decision of which experiments should be run.

BO employs a surrogate model, such as the ML models used in Chapter 2, to approximate the objective function. Built on top of this surrogate model is an acquisition function (AF), which guides the selection of inputs most likely to give

optimal outputs (**Figure 4.1**). The AF not only predicts potential outcomes but also quantifies the uncertainty of these predictions. One of the most used AFs is Expected Improvement (EI), which balances exploitation (selecting inputs likely to improve the current result) with exploration (sampling areas with high uncertainty to discovery new optima).

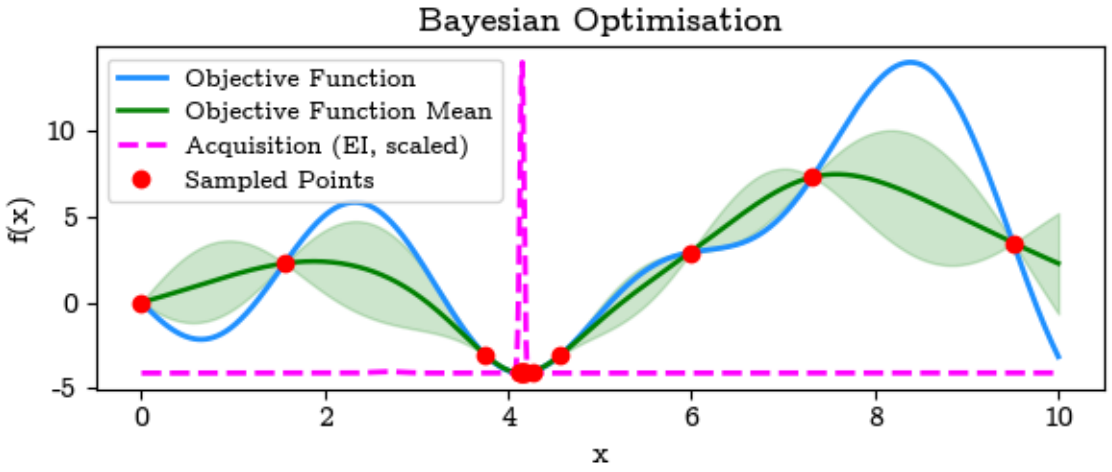


Figure 4.1: Illustration of BO applied to a one-dimensional objective function. The blue curve represents the optimisation process with its mean, and 95% confidence interval shown. The purple dashed line shows an acquisition function guiding the future selections

Through this method, BO can aid experimentation by iteratively training and retraining the surrogate model. At each step, the AF proposes new reaction conditions, until the optimal conditions are found.

4.1.3 Gaussian Process Regression

Gaussian Process Regression (GPR) is one of the most widely used surrogate models in BO, due to its flexibility and interpretability.[144, 275] Importantly, numerous ML studies identified GPR as the best-performing model, including in chemical property prediction.[276, 277]

GPR is non-parametric, meaning that it does not assume a specific functional form for the underlying data. Instead, it models the function as a collection of random variables with a joint Gaussian distribution (**Figure 4.2**). The relationship between inputs, the covariance, is described by the kernel function. The choice of kernel determines the smoothness, periodicity, and general shape of the functions that the

model can represent. One of the most used kernels is the Radial Basis Function (RBF) kernel, which assumes smooth, infinitely differentiable functions. The RBF kernel function is given by **Equation 4.1**.

$$k(X_i - X_j) = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) \qquad (4.1)$$

The kernel function quantifies the measure of the similarity between data points, the value of which depends on the distance between points, $X_i - X_j$, and the length-scale hyperparameter, σ. This length-scale hyperparameter is optimised to maximise the log-marginal likelihood of the model. High values of σ cause the kernel value to increase towards 1 and a low σ cause the kernel value to decrease towards 0. As such, at the two extremes of the hyperparameter σ of ∞ and 0, the data points will be treated as completely identical or completely different, respectively.[278]

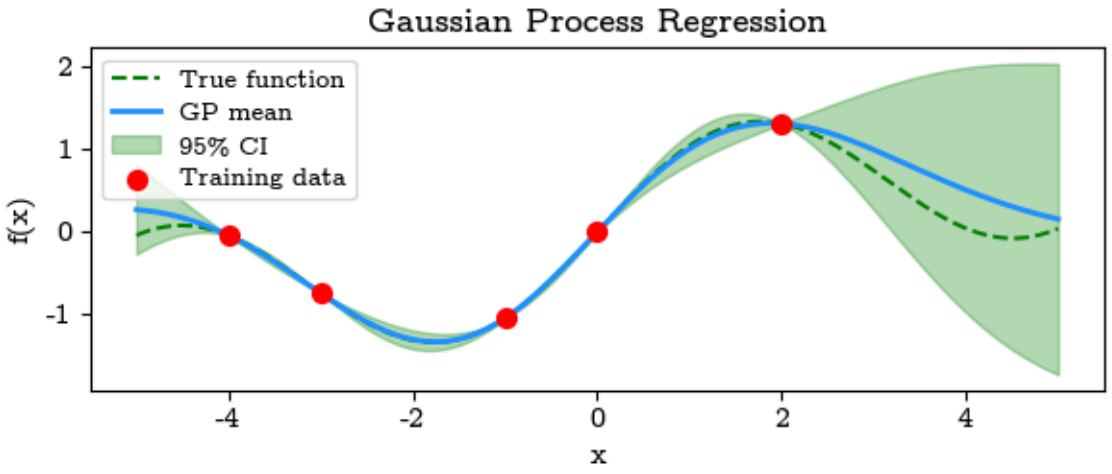


Figure 4.2: Depiction of GPR, the green dashed line represents the true function, the red points are the observed training points, while the blue line shows the GPR prediction.

4.1.4 Chemical Representations for Bayesian Optimisation

Just as with typical ML methods, a systematic representation of chemical information is required for input into BO frameworks. As optimisation campaigns are carried out in low-data regimes, where few reactions of this class exist, the choice of chemical representation can be challenging. The use of complex chemical descriptors on a low

number of data points can result in poor performance.[98] As such, this scenario may be better suited to the use of one-hot encoding (OHE).

OHE transforms categorical variables into distinct binary vectors. For example, reaction components are transformed into binary representations indicating whether a particular catalyst, solvent or additive from the design space is present in the reaction. OHE contains no explicit chemical or structural information but has demonstrated strong performance in ML applications. Notably, models using these simple encodings have, in some cases, achieved performances on par with those utilising computationally expensive QM descriptors.[279]

Such QM descriptors require mechanistic understanding, and have been used often for reaction optimisation in a synthetic setting, where the datasets used are much smaller. However, reports have suggested that these descriptors may contain little relevant information and add unnecessary noise to datasets.[133]

A report from the Gaunt, Lapkin and coworkers compared the use of different chemical representations on an iterative ML workflow on HTE dataset for prediction of reaction yields.[133] OHE was found to outperform complex descriptors incorporating chemical information, even in low data-regimes. Additionally, the use of more complex descriptors was found to require a larger initialisation dataset.

One technique used to retain more chemical information than OHE is the use of dimensionality reduction techniques coupled with the use of fingerprints. A commonly used example of this is principal component analysis (PCA). PCA is a dimensionality reduction technique able to retain the variance and the global structure of the data, with fewer dimensions. The high dimensional correlated variables are transformed to a smaller set of uncorrelated variables, known as principal components. Principal component 1 (PC1) contains the highest variance. In a two-dimensional plot, this would be the line of best fit. Principal component 2 (PC2) contains the next highest variance but is uncorrelated to PC1.[280] This continues for the later principal components. PCA is depicted in **Figure 4.3**.

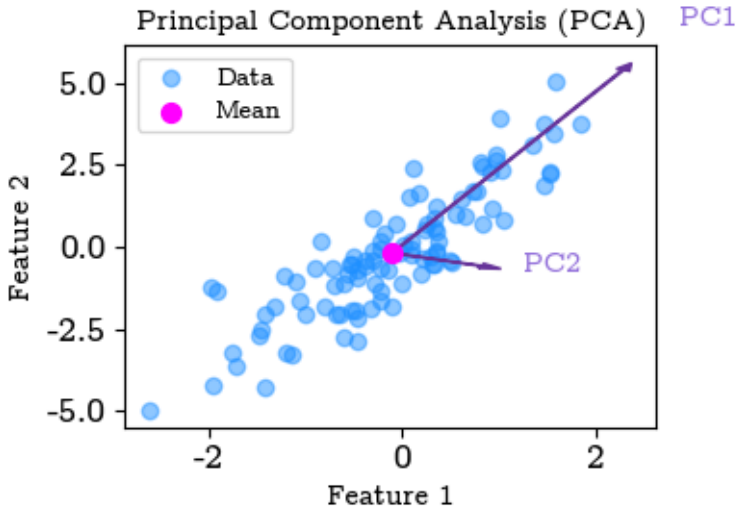


Figure 4.3: Illustration of PCA on a two-dimensional dataset.

4.1.5 Bayesian Optimisation for Reaction Optimisation

The conventional approach followed by synthetic chemists in an optimisation campaign typically involves the time-intensive screening of a relatively small subset of the vast space of possible reaction conditions. This process requires experts to simultaneously consider multiple reaction parameters such as solvent, temperature, pressure, catalyst loading and additives. The number of combinations increases exponentially with each additional parameter. Additionally, this strategy relies heavily on expert chemical intuition, mechanistic understanding and comprehensive literature knowledge to guide experimental choices. This often results in selection bias by proposing reaction conditions seen to work for similar reactions, but that may not be the optimal choice.[98]

Advances in HTE have allowed for the supplementation and replacement of manual screening in optimisation campaigns.[68] By facilitating exhaustive screening, HTE can reduce the reliance on expert-driven decision-making. However, access to HTE platforms remain limited, and these systems are not always compatible with reactions requiring harsh or inert conditions.[270, 281] Ideally, alternative approaches that minimise the number of experiments required to efficiently converge to optimal reaction conditions are preferable.

BO can be used to emulate, rather than replace, the decision-making process traditionally guided by domain experts, and has demonstrated success in ML driven navigation of reaction space.[273] Unlike human-guided approaches, BO is less susceptible to cognitive bias and can identify less obvious, yet potentially more advantageous, reaction conditions in a more data-efficient manner.

BO has attracted growing interest within the chemical community and has been applied to reaction optimisation strategies, most commonly for the use of automated flow chemistry.[78, 270, 271, 282] Notable BO frameworks such as Phoenics and Gryffin have been developed and employed in chemical case studies, including the identification of ligands and reaction conditions for Suzuki–Miyaura cross-coupling reactions, among other applications.[268, 283]

Many previous reports of implementing BO in chemistry have not been applicable to typical batch-scale screening, without the use of HTE equipment. Once such approach which has is the work from Doyle *et al,* who recently published an Experimental Design via Bayesian Optimisation framework (EDBO) compatible for human in-the-loop experimentation.[273] In this work, a BO framework was developed using experimental data from palladium-catalysed cross-coupling reactions, specifically the Suzuki–Miyaura and Buchwald–Hartwig reactions, with the objective of optimising reaction yield.[284] The framework employed a gaussian process surrogate model in conjunction with the parallel expected improvement AF. Various chemical representative features were explored, but reaction encodings derived from DFT calculations for the reaction components were selected. Initial results demonstrated that this BO framework outperformed traditional design of experiments (DOE) methods in identifying high-yielding conditions.[285]

The framework was subsequently tested on a novel palladium-catalysed C–H functionalisation reaction, where it was benchmarked against human chemists in identifying the global optimum yield within a reaction space of 1,728 possible conditions. Prior to the BO testing, all reactions within this space were run using HTE, allowing for an objective evaluation of performance. 50 chemists participated in a gamified challenge to locate the highest-yielding conditions, each working under an

experimental budget of 100 reactions, conducted over 20 batches of five reactions. Participants received real experimental outcomes after each batch to inform their next set of decisions. The BO framework consistently outperformed all human participants, identifying reaction conditions that achieved >99% yield in every run and discovered optimal conditions distinct from commonly used standards.

The optimiser was further evaluated on a significantly larger reaction space comprising 180,000 potential conditions for the Mitsunobu reaction.[286] In this case, the BO framework found a higher yield than the reaction control, which consisted of the typically reported parameters in the literature.

The work by Doyle *et al.* has displayed BO is a viable approach for reaction optimisation. To further enhance its applicability within the routine workflow of synthetic chemists, the implementation of real-time BO, without the reliance of HTE or QM calculations, would represent another step forward in making these methods more accessible and broadly adoptable in everyday laboratory practice.

4.1.6 Multi-objective Bayesian Optimisation

Although BO has been used successfully in locating reaction conditions that optimise one objective such as yield, reports of multi-objective optimisation remain scarce. Where explored, these studies often target objectives that are better suited to process chemistry, or do not accurately represent the complexities of synthetic chemistry.[287]

Straightforward screening methods for identifying optimal reaction conditions often require many objective evaluations and scale exponentially with the number of objectives involved. When multiple targets must be satisfied simultaneously, these methods become inefficient and may still fail to find ideal conditions. As a result, the time required becomes prohibitively high, which emphasizes the need for more efficient approaches. To address this, multi-objective optimisation algorithms are a more favourable approach.

To be used effectively in the context for reaction optimisation, BO must account for such multi-objective functions, such as yield and selectivity, simultaneously. Traditional single-objective ML approaches fail to capture the trade-offs between these

often-conflicting performance criteria. It is most likely that ideal conditions for which all outcomes assume their optimal values do not exist.

This concept is often described by the Pareto front, a formal concept implemented in multi-objective optimisation. A solution is on the Pareto front if no other solution is better for all objectives (**Figure 4.4**).[271, 288] These solutions are termed known dominating solutions, where one of the objectives cannot be improved without having a detrimental effect on the other objective.

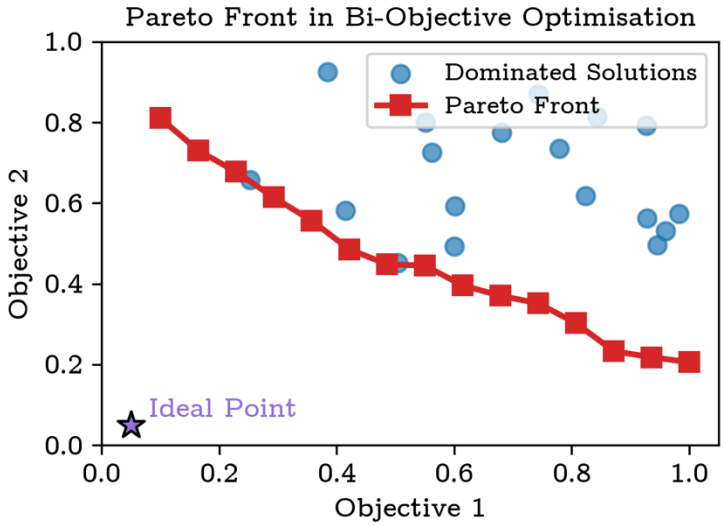


Figure 4.4: An example of a system with two competing objective functions. Where, it is impossible to reach the ideal point, where both objectives 1 and 2 are at their optimal values. Blue points correspond to dominated solutions that are outperformed by others. The red curve represents the Pareto front, the set of non-dominating solutions that define the optimal trade-off between objectives. The star marks the ideal point of an unattainable solution that maximises both objectives.

Accounting for the Pareto front and therefore generating optimisation algorithms that effectively capture the challenges of day-to-day research, would be more compelling for adoption in a synthetic chemist's reaction optimisation workflow. However, such ML workflows remain relatively scarce.

Several approaches have been proposed to address the challenges of multi-objective reaction optimisation. For instance, merit-based functions coupled with single-objective optimisation algorithms have been used to determine trade-offs between reaction rate and yield in the methylation of esters, as well as to balance rate and conversion in Paal–Knorr reactions.[289, 290] In another study focused on the synthesis of *o*-xylenyl adducts of buckminsterfullerene, a framework was implemented

in which a primary objective was optimised while secondary objectives were maintained within specified constraints.[291]

More recently, Doyle *et al.* extended their EDBO framework (developing EDBO$^+$) to a multi-objective setting for the optimisation of both yield and *ee* in a Ni/photoredox-catalysed enantioselective cross-coupling reaction. They employed the q-Expected Hypervolume Improvement (q-EHVI) acquisition function, which estimates the expected gain in Pareto hypervolume, to efficiently navigate a reaction space of 1,728 possible conditions and identify optimal trade-offs.[292]

A method reported by Amar *et al.*, used the Thompson Sampling Efficient Multi-Objective Optimisation (TS-EMO) algorithm.[293] The TS-EMO algorithm was effective at approximating experimental conditions that lie on the Pareto front of yield and diastereoseletivity of an AH reaction.[294] In this work, GPR surrogate models were trained on an experimental dataset of dicarbonyl(acetylacetonato)rhodium(I) ($Rh(CO)_2(acac)$)/Josiphos catalysed AH of a chiral α,β- unsaturated δ-lactam. The approach was successful in locating over 15 solvents that achieved over 90% conversion, where only one of these solvents existed in their training set that was suggested by human experts.

An alternative approach to multi-objective optimisation is Chimera, a hierarchy-based scalarization function.[295] In this method, objectives are prioritised by assigning a hierarchy, with more critical objectives ranked higher. Each objective is also associated with a tolerance, defining the acceptable deviation from its ideal value. Optimisation proceeds by considering the higher objectives first; lower-priority objectives are only optimised once all higher-priority objectives have met their tolerance thresholds. Additionally, improvements to lower-tier objectives must not compromise higher-tier objectives beyond their specified tolerances.

Chimera has been implemented successfully in chemical applications, notably in the Gryffin optimisation framework, where it was applied to the multi-objective optimisation of a Suzuki–Miyaura cross-coupling reaction.[283]

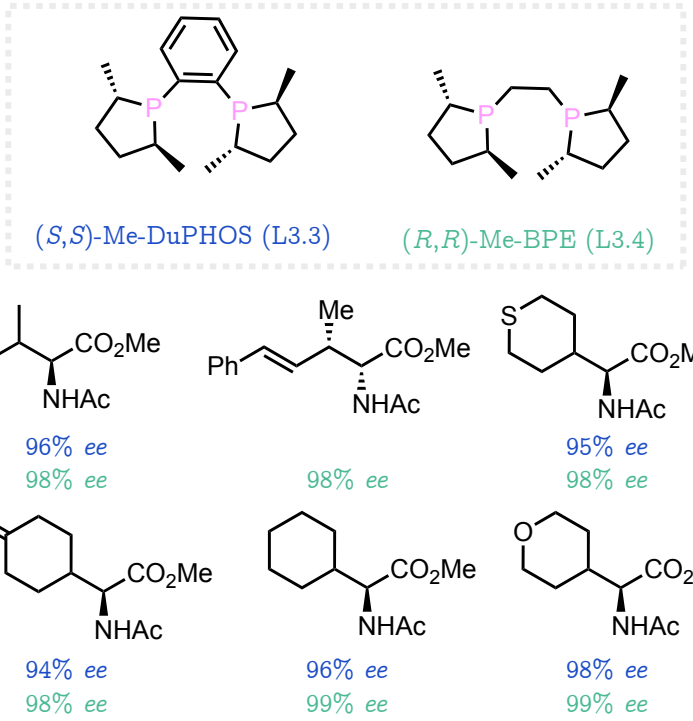
4.1.7 Asymmetric Hydrogenation of Tetra-substituted Enamides

Enamides have attracted attention in the AH of TSC=Cs, due to their ability to provide direct access to a wide variety of chiral substituted amines and amides, which are valuable intermediates in the synthesis of natural products, biologically active compounds and drugs.

As discussed in Chapter 1, the AH of this class of compounds has been much more successful for the di- and tri-substituted enamides, while tetra-substituted enamides continue to remain challenging. However, several reports have demonstrated successful results with good yields and selectivities.[29, 41, 42, 44, 203, 206, 222, 296-301]

A pivotal advancement was reported by Burk *et al.* in 1995, demonstrating the AH of both cyclic and acyclic tetra-substituted enamides with high *ee* using cationic diphosphine-rhodium catalysts.[41] This marked the first report for the AH of β,β-disubstituted α-(acylamino)acrylates and showed compatibility with both *E*- or *Z*-configured substrates. Their system employed chiral ligands Me-BPE and Me-DuPhos in MeOH under 90 psi of H_2 (**Scheme 4.1**).

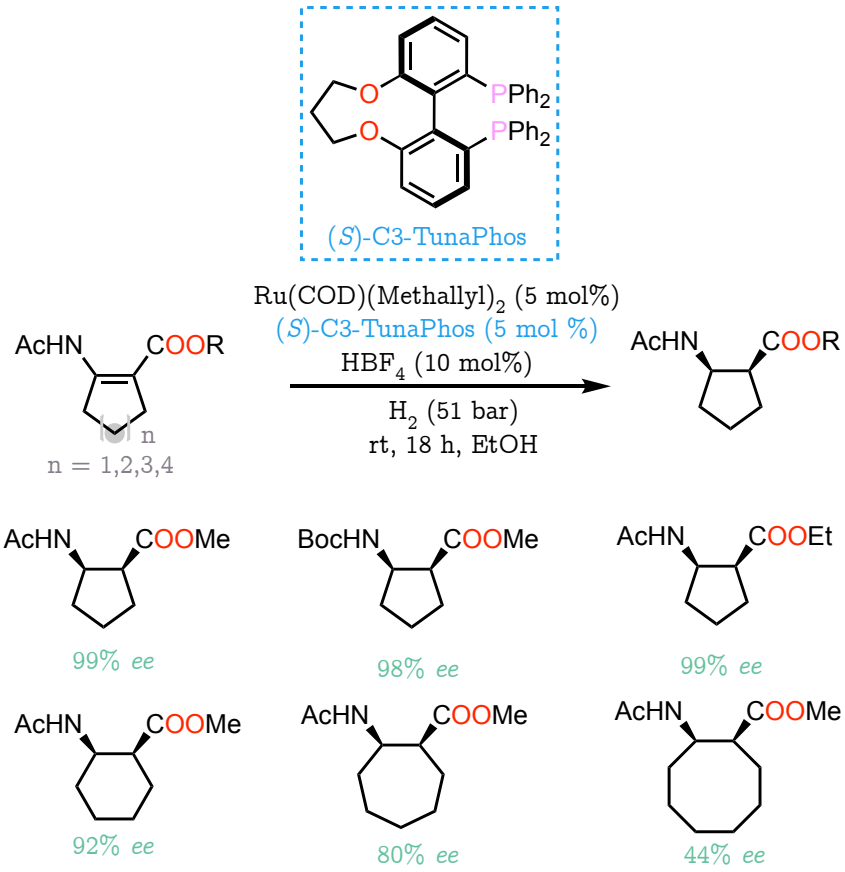
Burk (1995)



Scheme 4.1: Examples of the substrate scope reported by Burk *et al.* Reactions were conducted at 25 °C under 90 psi of H_2 for 12-24 h with 0.2 mol% catalyst precursors [(*S*,*S*)-Me-DuPHOS-Rh(COD)]$^+$OTf$^-$ or [(*R*,*R*)-Me-BPE-Rh(COD)]$^+$OTf$^-$.

In 2003, another advancement in the AH of cyclic enamides was reported by the Zhang group, with the enantioselective hydrogenation of cyclic β-(acylamino)acrylates applying a Ru-C3-TunaPos catalytic system. While the method exhibited high enantioselectivity for most cyclic substrates, a notable drop in selectivity was observed when the ring size was extended to seven- or eight-membered substrates (**Scheme 4.2**).[301]

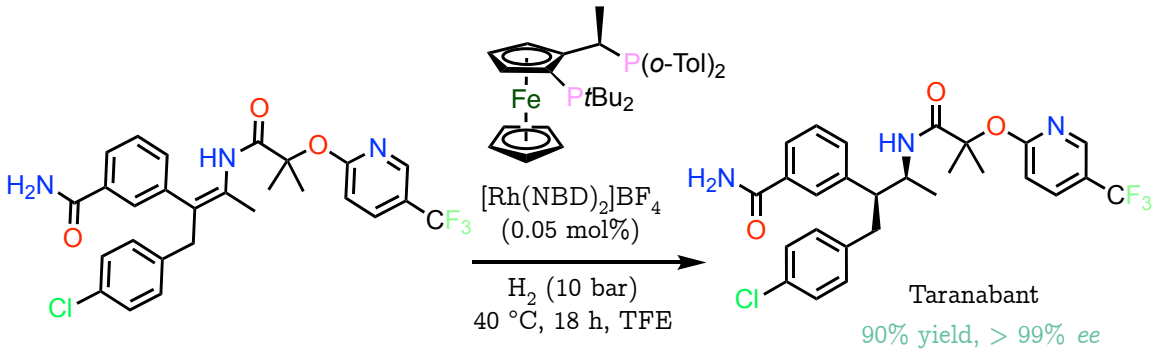
Zhang (2003)



Scheme 4.2: Examples of the AH of β-(acylamino)acrylates applying a Ru-(*S*)-C3-TunaPos catalyst.

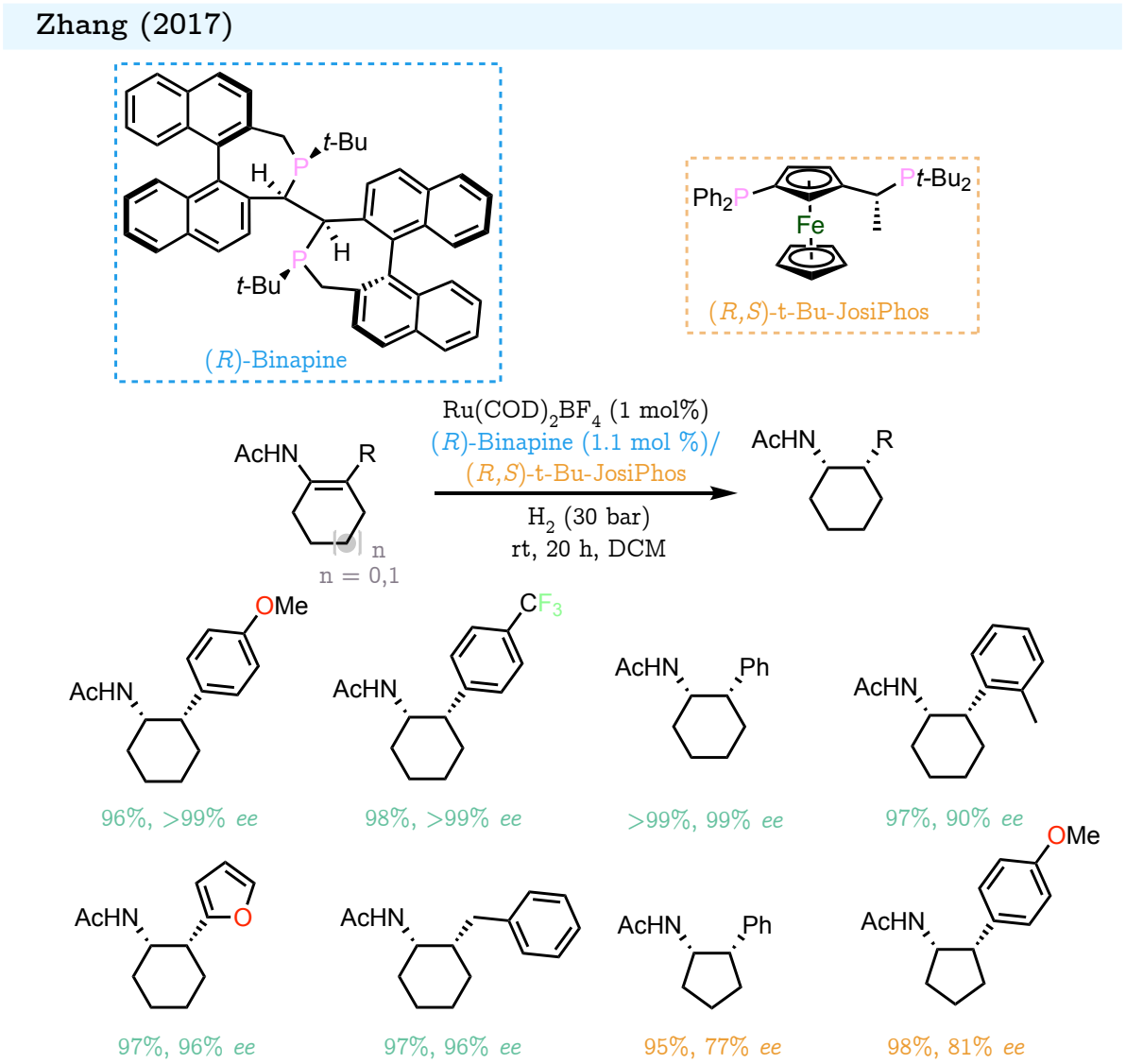
This chapter is focused on the AH of cyclic enamides. However, a noteworthy example involving an acyclic enamide is a 2009 study by a Merck team,[297] which demonstrated the critical role of AH of enamides in drug development through the synthesis of taranabant, a cannabinoid-1 receptor investigated for the treatment of obesity.[302] Chirality was induced in the final step of the synthesis using a chiral ferrocene-Rh catalyst (**Scheme 4.3**).

Merck (2009)



Scheme 4.3: The last step in the synthesis of Taranabant, developed by Merck.

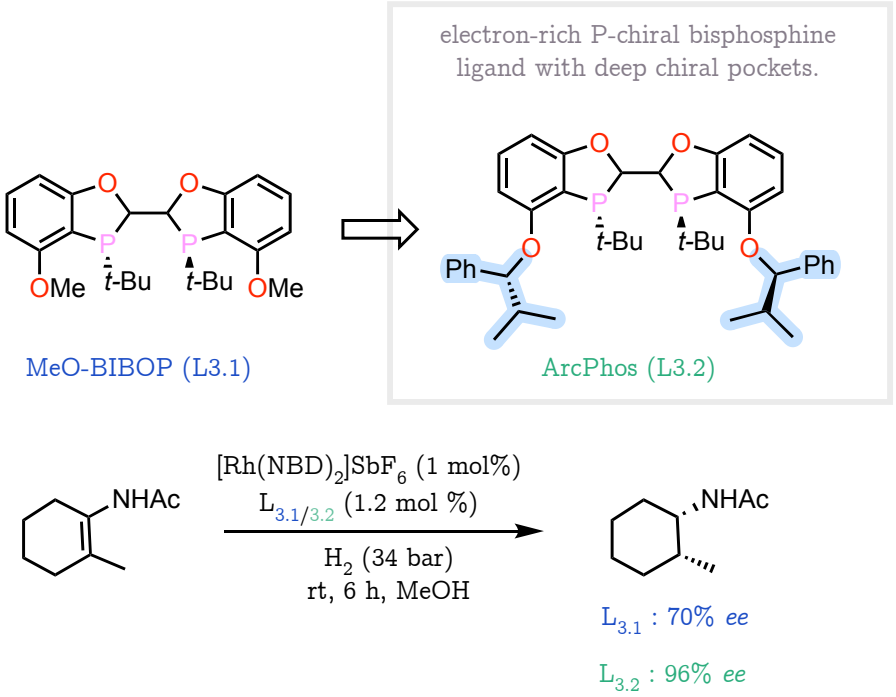
In 2017, the Zhang group reported the AH of β-aryl substituted tetra-substituted enamides applying a Rh-Binapine catalytic system (**Scheme 4.4**).[203] A range of carbocylic substrates exhibited high *ee* under mild condition. However, for the cyclopentene enamides a different ligand, (*R*,*S*)-t-Bu-JosiPhos was required, and a reduction in enantioselectivity was still observed for these substrates.



Scheme 4.4: Examples of the substrate scope reported by the Zhang group for the AH of β-aryl substituted tetra-substituted enamides, showing corresponding conversion and *ee*.

In 2019, the Tang group, in collaboration with Pfizer, published the AH of a range carbocylic and heterocyclic tetra-substituted enamides, using a rhodium catalyst bearing a new chiral bisphopshine ligand (**Scheme 4.5**).[4]

Tang (2019)

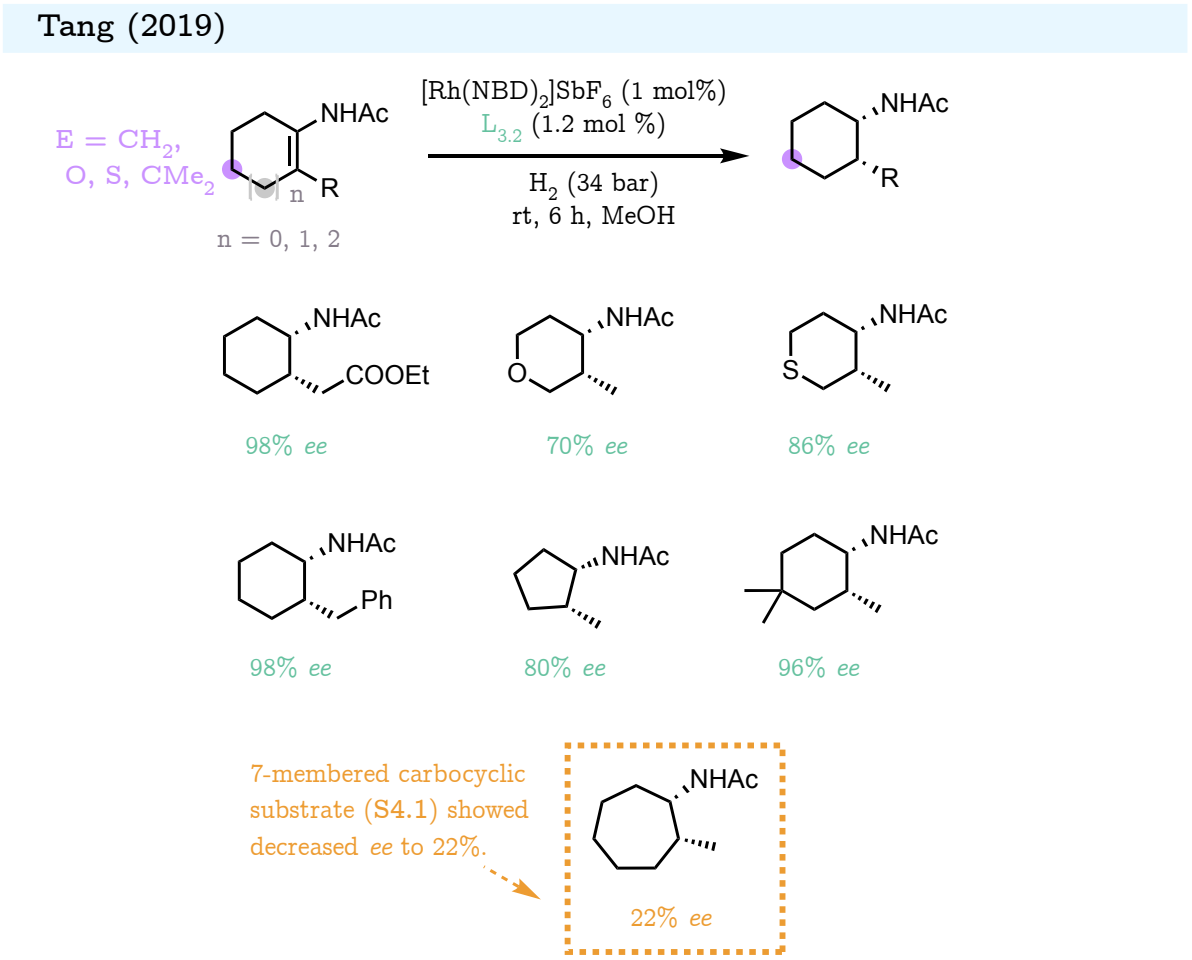


Scheme 4.5: Development of the ArcPhos ligand for the AH of N-(2-methylcyclohex-1-en-1-yl)acetamide.

.

Using the model substrate, *N*-(2-methylcyclohex-1-en-1-yl)acetamide, it was found the commercially available chiral ligand, (*R*,*R*,*R*,*R*,)-MeO-BIPOP showed adequate enantioselectivity of 70% *ee*. Through modifying the 4,4'-positions of MeO-BIPOP with α-isopropyl-substituted benzyloxy groups, the enantioselectivity increased to 96% *ee*, thus developing the new chiral bisphosphine ligand, ArcPhos.

This system was used to produce a range of carbocyclic and heterocyclic amines, with high yields and enantioselectivities, including compounds containing an ester functional group, as well as amino acid, thiopyran and pyran derivatives (**Scheme 4.6**). Surprisingly, the carbocyclic amide, *N*-(-2-methylcycloheptyl)acetamide (**S4.1**), bearing a 7-membered ring, showed a poor *ee* of 22%.



Scheme 4.6: The substrate scope of heterocylic and carbocyclic amines published by the Tang group. Displaying the low *ee* of 22% from substrates S4.1.

The system was used for the highly enantioselective synthesis of (*R*)-Tofacitinib, a Janus kinase (JAK) inhibitor used for the treatment of rheumatoid arthritis.[9] Previously, the synthesis of (*R*)-Tofacitinib relied on the kinetic resolution of its racemic mixture, resulting in significant waste. This AH system developed by the Tang group has offered a more efficient alternative, substantially reducing waste in this synthesis as well as in the production of future drug compounds containing this structural motif.

Although the reaction conditions developed for this subset of enamides yielded excellent results for several compounds, the poor performance observed with substrate **S4.1** suggests that the system, and the ligand ArcPhos in particular, was not well-suited for this substrate. Furthermore, to date, no alternative systems have been reported for the synthesis of this compound.

Previous studies on the AH of enamides have highlighted the strong substrate–ligand dependence of this reaction. Even minor structural changes to the substrate can lead to significant drops in enantioselectivity, and in some cases, entirely different ligands are required to achieve high *ee*. Given the low selectivity observed for **S4.1** with the reported system, the development of a new catalytic system is required for the enantioselective synthesis of pharmaceutically relevant compounds containing this 7-membered ring motif. This would enable broader access to structurally diverse chiral amide intermediates with potential applications in medicinal chemistry.

4.1.8 Chapter Aims

This chapter aimed to optimise the AH reaction conditions for *N*-(-2-methylcycloheptyl)acetamide (**S4.1**), using a BO approach. The objective was to identify the most enantioselective and high-yielding reaction from a predefined set of reactions, employing a more data-efficient strategy than a traditional screening approach. To quantify achievement of this goal, an experimental budget of 10% of the reactions in the predefined set was employed. To ensure compatibility with existing synthetic practices for this reaction class, the BO was implemented in real time, guided solely by experimental results from reactions selected by the model, without reliance on HTE or QM calculations.

4.2 METHODS

The revised approach to predict the most successful AH for **S4.1** begun by defining a diverse space of distinct reaction conditions. From this space, input features were generated that are proficient for use in an ML model to guide *ee* prediction. Initially, a set of reactions were carried out, and their experimental *ee* values and conversion to product measured to train the surrogate model. The AF then measured the EI scores of all the reactions in the set. The reactions with the highest EI score were then selected, performed experimentally, and their *ee* values and conversion to product measured. The new data points were then added to the training set, and the surrogate model was retrained. This iterative process was repeated until the most selective reaction was identified, as illustrated in **Figure 4.5**.

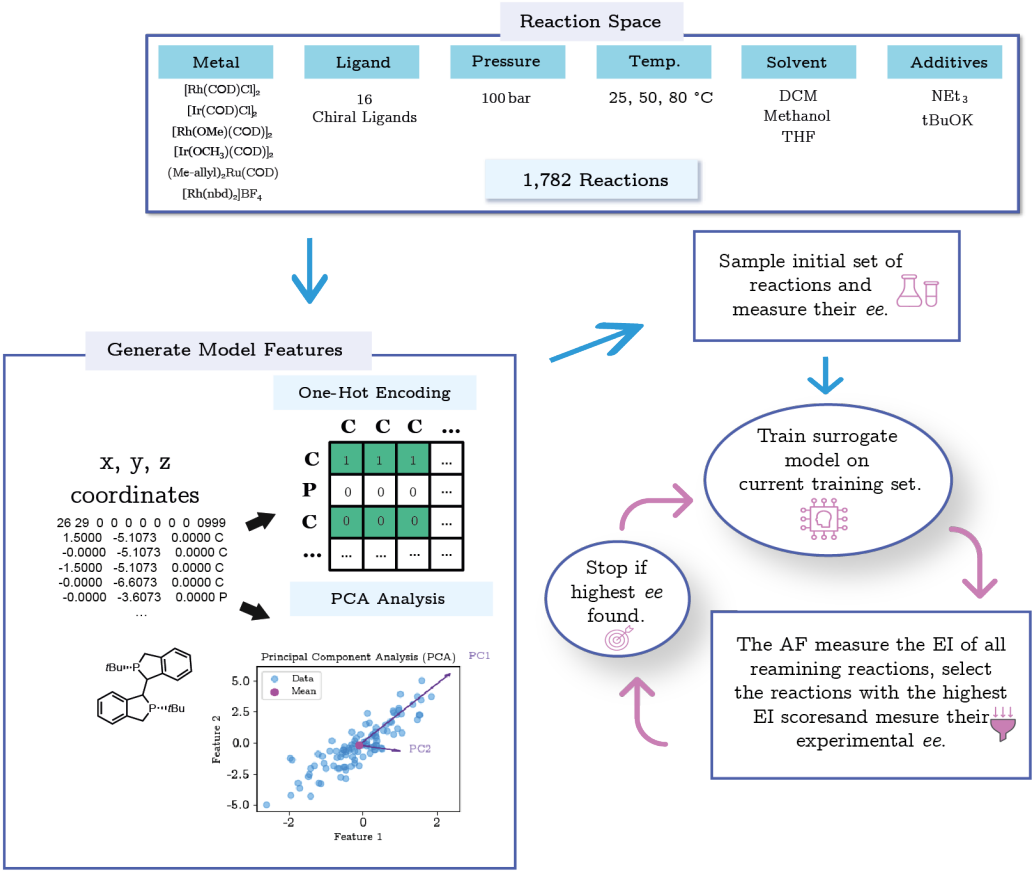


Figure 4.5: Workflow for BO approach for reaction conditions. The reaction space is defined by varying metals, ligands, solvents, additives, pressures, and temperatures, resulting in 1,782 possible reactions. Features are generated to represent each reaction. An initial subset of reactions is sampled and *ee* and conversion measured, after which a GPR surrogate model is trained. The AF evaluates the remaining candidates, selecting those with the highest EI for experimental validation. This loop is repeated until the optimal *ee* is identified.

4.2.1 Reaction Space

The prespecified set of reactions were designed by selecting the most used reaction conditions, while also exploiting the use of commercially available chiral ligands. A total of six pre-catalysts were chosen: chloro(1,5-cyclooctadiene)rhodium(I) dimer ($[Rh(COD)Cl]_2$), chloro(1,5-cyclooctadiene)iridium(I) dimer ($[Ir(COD)Cl]_2$), bis(μ-methoxo)bis(1,5-cyclooctadiene)dirhodium(I) ($[Rh(OMe)(COD)]_2$), bis(μ-methoxo)bis(1,5-cyclooctadiene)diiridium(I) ($[Ir(OMe)(COD)]_2$), bis(μ^3-2-methylallyl)(1,5-cyclooctadiene)ruthenium(II) ($(Me\text{-}allyl)_2Ru(COD)$) and $[Rh(NBD)_2]BF_4$. 16 chiral ligands were chosen and their structures are shown in **Figure 4.6**. The three most prevalent solvents in the literature were used: DCM, MeOH and THF. Three temperatures used were: 25, 50 and 80 °C. As well as two additives seen previous success in literature. Triethylamine (NEt_3) and *t*BuOK. These conditions defined a reaction space comprising 1,782 distinct reactions (**Figure 4.6**).

The remaining conditions: pressure (100 bar), time (24 h), pre-catalyst (5 mol%), chiral ligand (5.75 mol%) and additive loading (20 mol%) were all kept constant.

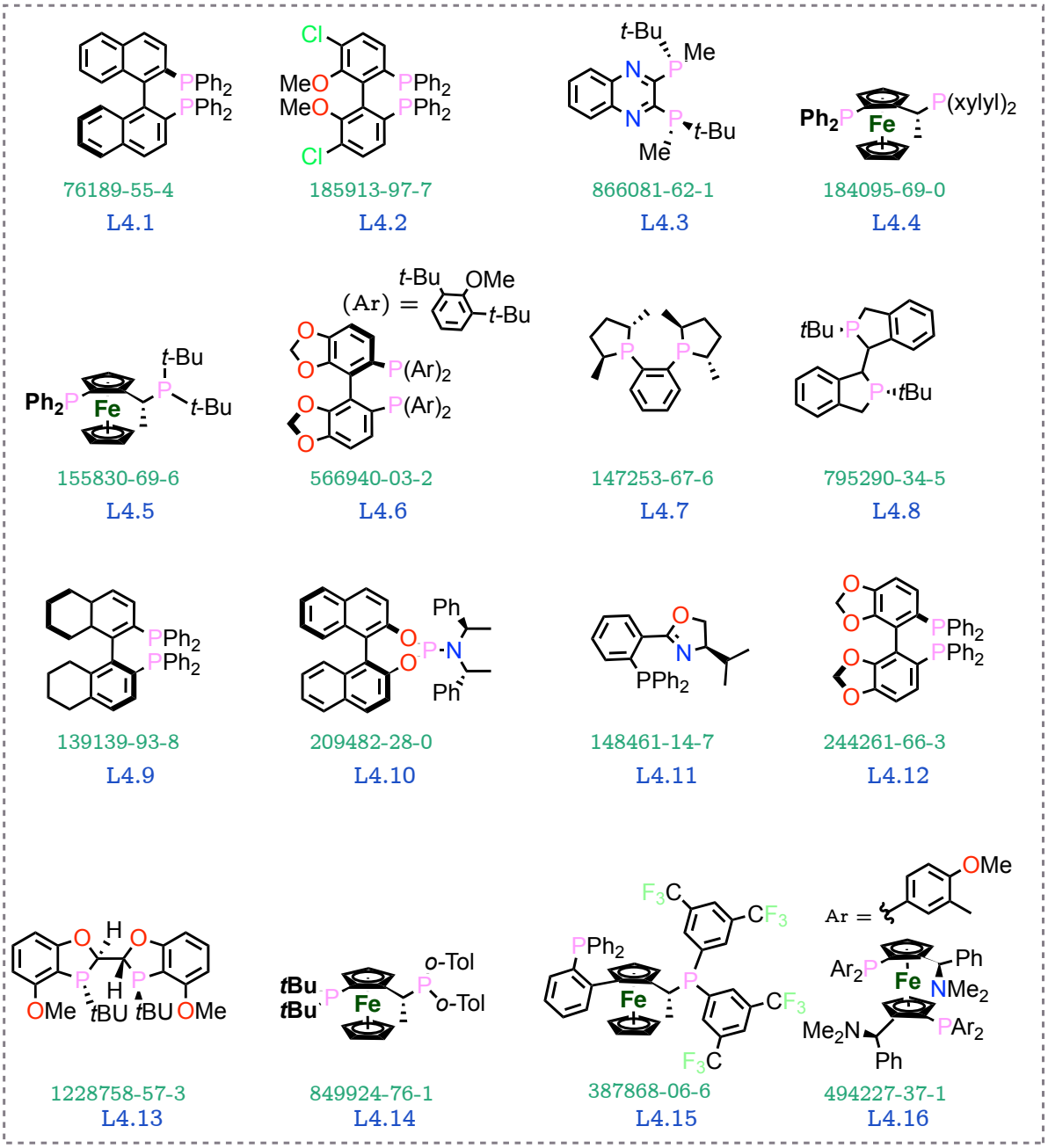


Figure 4.6: The reaction space used in this project, comprised of 16 chiral ligands, shown with their corresponding CAS number, 3 solvents, 3 temperatures and 2 additives.

4.2.2 Model Features

Several different model features were considered, specifically one-hot encoding, due to its success seen in previous publications. All reaction variables including pre-catalysts, solvents, temperature, and additives were all represented as one-hot encoded features. However, due to the increased complexity of the chiral ligands used in comparison to the other features, they were represented using PCA of their respective ECFPs.

In this work, four principal components were used. PCA was performed on the ECFPs for all the chiral ligands created from the ligand co-ordinates by RDKit, with radius three and number of bits set to 4096. All bits that were zero across all ligands were removed, with PCA performed on the remaining bits.

4.2.3 Initial Reactions

The initial set of reactions were selected to ensure each variable reaction condition was represented approximately equally. 48 initial reactions were conducted as the input for BO, 12 reactions were conducted in each solvent, 8 with each additive, 24 at RT, 12 at 50 and 12 at 80 °C. Each ligand was used at least once.

4.2.4 Machine Learning Methods

For the BO, the surrogate model used was GPR, with the RBF kernel, and a white kernal with noise level set to 0.01, implemented by the sklearn python package.[303] The AF used was EI.

Prior to model training, the hyperparameters for the GPR model with the RBF kernel were tuned, specifically the value of σ. This was carried out initially and remained constant during the BO iterations. This is unlike the default setting for the GPR model in sklearn, which automatically optimises the hyperparameters of the kernel function by maximising the log-marginal likelihood of the model with respect to the hyperparameters. This was instead kept constant as tuning the hyperparameter to small amounts of training data can cause the model to overfit, resulting in poor model performance.[304] This method was implemented following the approach described by Tang *et al.*[305] The RBF kernel parameter σ was selected prior to the collection of

any training data. The value was selected to maximise the variance of the data in the kernel space, thereby encapsulate the diversity of the dataset. The goal was to choose a value of σ that maximises the variance computed across all distinct pairs of data points, enhancing the model's ability to distinguish between them in kernel space.

This variance is a mono-model function of σ, meaning it has a single peak. To identify the optimal value, a binary search was performed over a grid of candidate σ values, stopping when the value that produced the maximum variance was found (**Figure 4.7**).

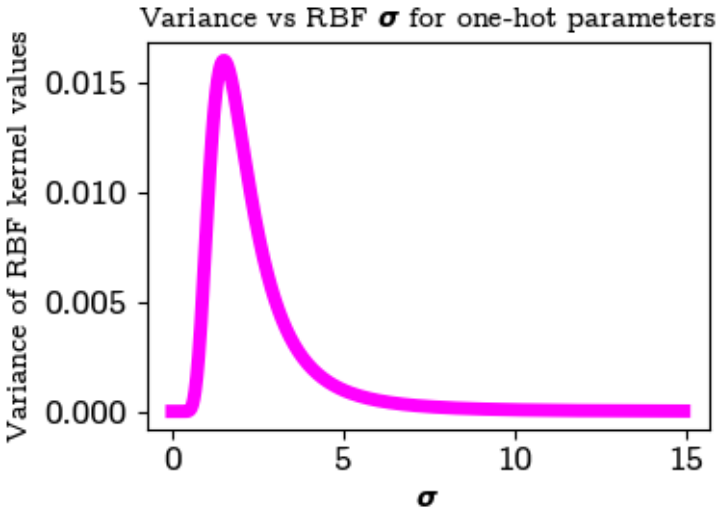


Figure 4.7: Plot of variance of how the variance of the RBF kernel space changes with the hyperparameter σ.

4.2.5 Conversion Tolerance

Although the primary goal of this project was to optimize for enantioselectivity, achieving sufficient product formation is a prerequisite. Therefore, when the reaction conversion was low, maximizing conversion took priority. A conversion threshold of 10% was implemented in the BO approach using the Chimera Hierarchy-based Scalarizing Function.[295] Specifically, when conversion was below 10%, the algorithm prioritized conversion as the main objective, once conversion exceeded 10%, the algorithm shifted to optimizing enantioselectivity.

The overall values for the objective function used to train the surrogate GPR model during BO is given by **Equation 4.2**.

$$Obj_{train} = ee_{train} \cdot \Theta^{-}(yield_{train}, tol, damp) + yield_{train} \cdot \Theta^{+}(yield_{train}, tol, damp) \quad (4.2)$$

Where:

$$\Theta^{-}(y, t, d) = \frac{1}{1 - \exp\left(-\frac{(y - t)}{d}\right)}$$

$$\Theta^{+}(y, t, d) = 1 - \Theta^{-}(y, t, d)$$

Thus, when conversion is low (below tolerance of 10%), the objective function values are smoothly scaled to become the values of conversion, and when the conversion is high the objective values become the enantioselectivities. A model of this is given in **Figure 4.8**, where the extreme examples are shown. This figure illustrates that if the conversion tolerance was set to 0% the objective function would optimise purely for enantioselectivity and when the tolerance is set to 100% the objective function would optimise purely for conversion to product. The bottom graph in this picture illustrates the tolerance of 10% used in this project.

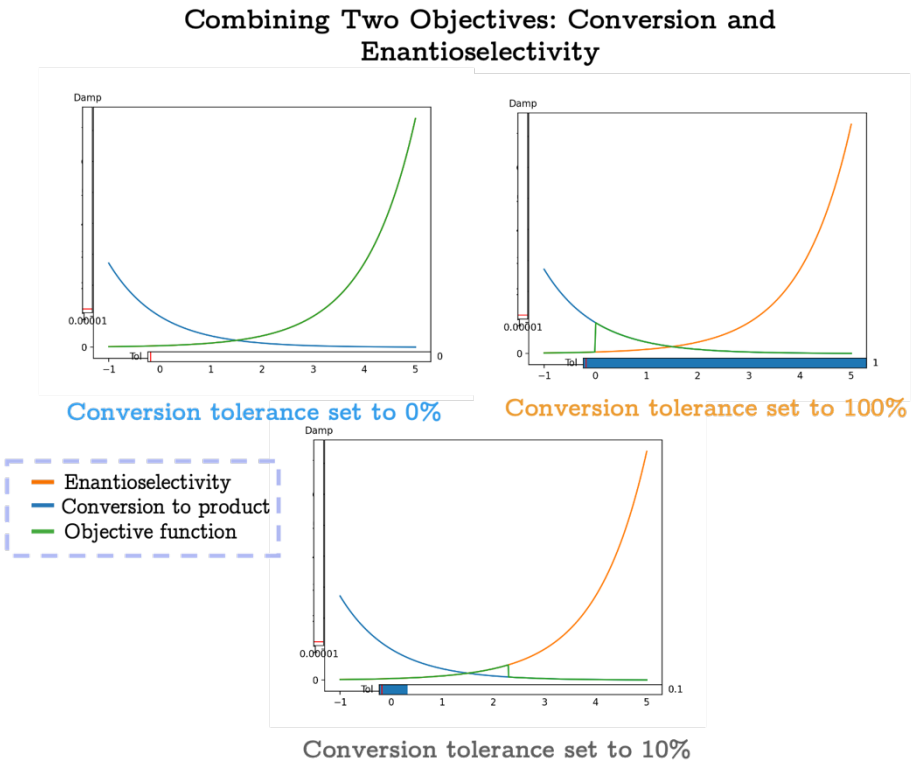


Figure 4.8: Visualisation of the objective function combines the two objectives. In this model example, the orange curve represents enantioselectivity, the blue curve represents conversion to product and the green curve represents the combination of the two objective functions.

4.2.6 Batch Selection

To select the next batch of reactions all the remaining reactions from the reaction space not currently in the training set were ranked by their EI scores and the reactions with the highest scores were selected.

As the reactions were carried out in a HEL cat24 autoclave, capable of carrying out 24 reactions at once, the number of reactions selected for each batch was 24 to maximise the number of data points obtained in a smaller time frame. However, as the model was only 'learning' every batch this was later reduced to 12 to allow the model a higher number of opportunities to 'learn', this will be discussed further in **Section 4.3**. All reactions in each batch were required to be conducted at the same temperature. Therefore, the temperature for the entire batch was set according to the specified temperature of the reaction with the highest EI score.

4.3 RESULTS AND DISCUSSION

4.3.1 Principal Component Analysis

To capture more detailed chemical information than OHE, PCA was performed on ECFPs. This allowed for the retention of key structural variance while reducing the high dimensionality of the fingerprint space. Four principal components were computed, with the PCA projection highlighting variance beyond the dominant PC1. The distribution of ligands in this principal component space categorised by PC1-PC3 and PC2-PC4 is shown in **Figure 4.9**. Ligands that cluster closely in this plot are expected to share similar functional features, while those far apart represent more distinct structural chemotypes. The right plot PC2-PC4 highlights the variation orthogonal to PC1, therefore revealing subtle structural differences not captured in the first plot. Visualising the ligands in this reduced-dimensional space shows that the library is not biased toward any specific region, indicating a diverse input space for initiating BO.

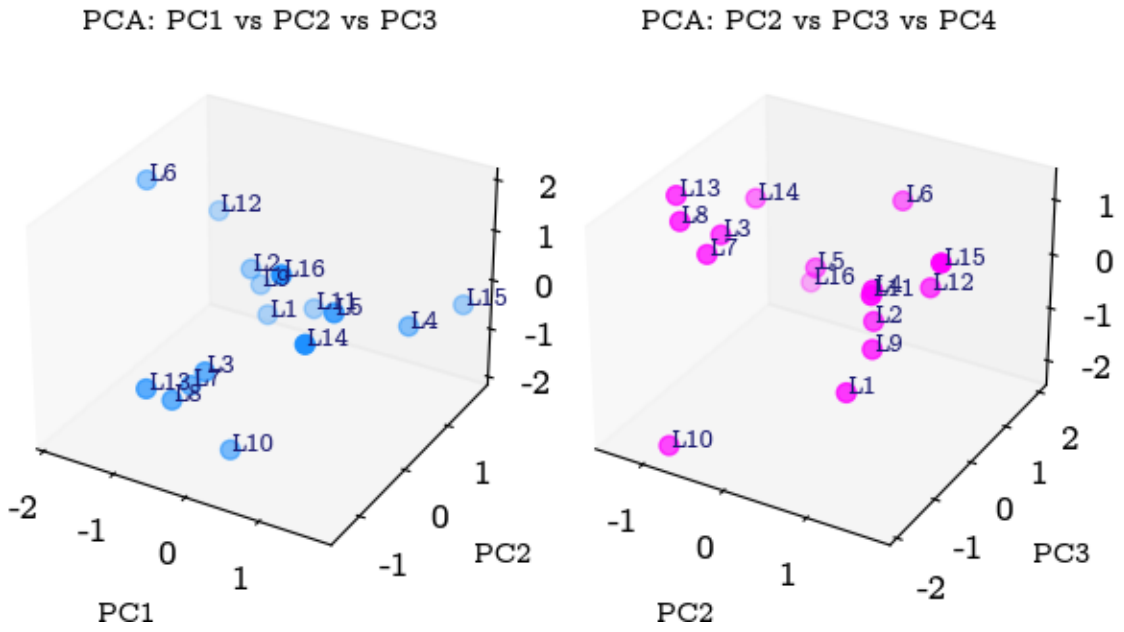


Figure 4.9: PCA of ligand molecular fingerprints, showing the distribution of ligands in the PC1-PC2-PC3 (blue) and PC2–PC3–PC4 (purple) space.

4.3.2 Initial Set of Reactions

The initial set of reactions used to begin the BO workflow are shown in **Table 4.1**. The reactions showed promising results, with several reactions achieving higher *ee* values than those reported in the literature. However, enantioselectivity remained low in many cases: 26 of the 48 reactions showed either an *ee* below 5% or conversion too low to determine *ee*. Of the reactions with measurable conversion, ten had *ee* values below 5%. The mean *ee* across all reactions was 13.6%, excluding those for which *ee* could not be determined. Notably, 17 reactions reached full conversion, while 13 exhibited conversions below 5%, supporting the decision to apply a 10% conversion threshold for the BO. Although no clear visual trend was observed between *ee* and conversion, linear regression analysis indicated a slight positive correlation (**Figure 4.10**).

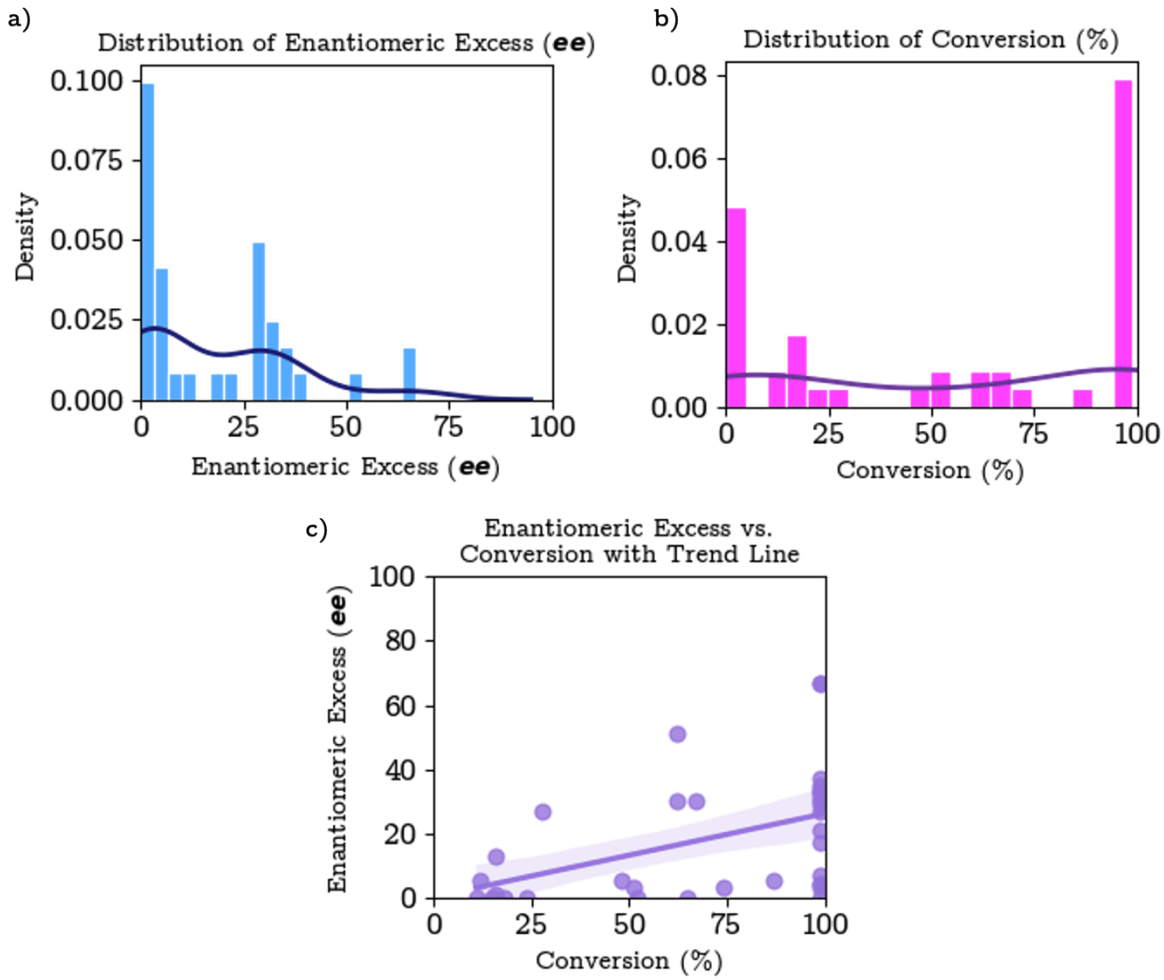


Figure 4.10: a) and b) show the distribution of *ee* and conversion across the initial reaction dataset. The histogram represents the frequency of the observed ee values, with the overlaid KDE shows a smoothed approximation of the distribution. c) Scatter plot showing the relationship between *ee* and conversion (%) for the initial reactions. A linear regression is fitted to the data, with the shaded region representing the 95% confidence interval.

Table 4.1: The 48 initial reactions conducted. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.1	**L4.1**	$Rh(COD)Cl]_2$	DCM	*t*BuOK	50	62	51
4.2	**L4.2**	$[Ir(COD)Cl]_2$	THF		50	28	27
4.3	**L4.3**	$[Rh(OMe)(COD)]_2$	DCM		50	48	5
4.4	**L4.4**	$[Ir(OMe)(COD)]_2$	THF		50	0	-
4.5	**L4.14**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	35
4.6	**L4.7**	$[Rh(NBD)_2]BF_4$	THF	*t*BuOK	50	100	3
4.7	**L4.1**	$Rh(COD)Cl]_2$	DCM		50	100	4
4.8	**L4.9**	$[Ir(COD)Cl]_2$	THF		50	0	-
4.9	**L4.5**	$[Rh(OMe)(COD)]_2$	DCM		50	99	67
4.10	**L4.15**	$[Ir(OMe)(COD)]_2$	THF		50	0	-
4.11	**L4.16**	$(Me\text{-}allyl)_2Ru(COD)$	DCM	*t*BuOK	50	100	7
4.12	**L4.8**	$[Rh(NBD)_2]BF_4$	THF		50	99	31
4.13	**L4.10**	$Rh(COD)Cl]_2$	DCM		rt	0	-
4.14	**L4.10**	$[Ir(COD)Cl]_2$	THF		rt	62	30
4.15	**L4.3**	$[Rh(OMe)(COD)]_2$	DCM	NEt_3	rt	29	3
4.16	**L4.12**	$[Ir(OMe)(COD)]_2$	THF		rt	0	-
4.17	**L4.13**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		rt	16	13
4.18	**L4.11**	$[Rh(NBD)_2]BF_4$	THF		rt	0	-
4.19	**L4.1**	$Rh(COD)Cl]_2$	DCM		rt	52	0
4.20	**L4.2**	$[Ir(COD)Cl]_2$	THF	NEt_3	rt	0	-
4.21	**L4.3**	$[Rh(OMe)(COD)]_2$	DCM		rt	0	-
4.22	**L4.4**	$[Ir(OMe)(COD)]_2$	THF		rt	0	-
4.23	**L4.13**	$[Rh(NBD)_2]BF_4$	DCM		rt	99	0
4.24	**L4.11**	$Rh(COD)Cl]_2$	THF		rt	65	0
4.25	**L4.16**	$[Ir(COD)Cl]_2$	MeOH	NEt_3	rt	24	0
4.26	**L4.10**	$[Rh(OMe)(COD)]_2$	MeOH		rt	15	0
4.27	**L4.11**	$[Ir(OMe)(COD)]_2$	MeOH		rt	0	-
4.28	**L4.15**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH		rt	12	5
4.29	**L4.10**	$[Rh(NBD)_2]BF_4$	MeOH		rt	51	3
4.30	**L4.12**	$Rh(COD)Cl]_2$	MeOH	NEt_3	rt	99	21
4.31	**L4.9**	$[Rh(NBD)_2]BF_4$	MeOH		rt	99	27
4.32	**L4.6**	$[Rh(OMe)(COD)]_2$	MeOH		rt	99	29
4.33	**L4.1**	$[Ir(OMe)(COD)]_2$	MeOH		rt	18	0
4.34	**L4.2**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH		rt	99	34
4.35	**L4.3**	$[Ir(COD)Cl]_2$	MeOH	NEt_3	rt	0	-
4.36	**L4.11**	$[Rh(NBD)_2]BF_4$	MeOH	*t*BuOK	rt	0	-
4.37	**L4.2**	$Rh(COD)Cl]_2$	THF	*t*BuOK	80	67	30
4.38	**L4.4**	$[Ir(COD)Cl]_2$	THF		80	74	3
4.39	**L4.6**	$[Ir(COD)Cl]_2$	THF	NEt_3	80	87	5
4.40	**L4.1**	$[Ir(OMe)(COD)]_2$	THF		80	16	1

4.41	**L4.9**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	99	37
4.42	**L4.5**	$Rh(NBD)_2]BF_4$	MeOH		80	99	67
4.43	**L4.16**	$Rh(OMe)(COD)]_2$	MeOH	NEt_3	80	99	17
4.44	**L4.8**	$[Rh(COD)Cl]_2$	MeOH		80	99	33
4.45	**L4.10**	$[Ir(OMe)(COD)]_2$	DCM	*t*BuOK	80	11	0
4.46	**L4.10**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		80	99	30
4.47	**L4.3**	$[Rh(NBD)_2]BF_4$	DCM		80	99	4
4.48	**L4.12**	$Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	99	33

Figure 4.11 and **Figure 4.12** show the distribution of *ee* across the different pre-catalyst and ligands evaluated. Overall, the performance among the pre-catalysts was relatively consistent, with no single complex demonstrating a markedly superior outcome. Notably, $[Ir(COD)Cl]_2$, delivered moderate enantioselectivities of 27 and 30%, with ligands **L4.2** and **L4.10** respectively (**Scheme 4.7**). This is particularly interesting given that iridium-based pre-catalysts have been scarcely explored in the literature for this class of substrates. However, $[Ir(OMe)(COD)]_2$ exhibited notably poor performance, yielding low *ee* across all instances and failing to produce any measurable product conversion in six out of the eight reactions in which it was employed.

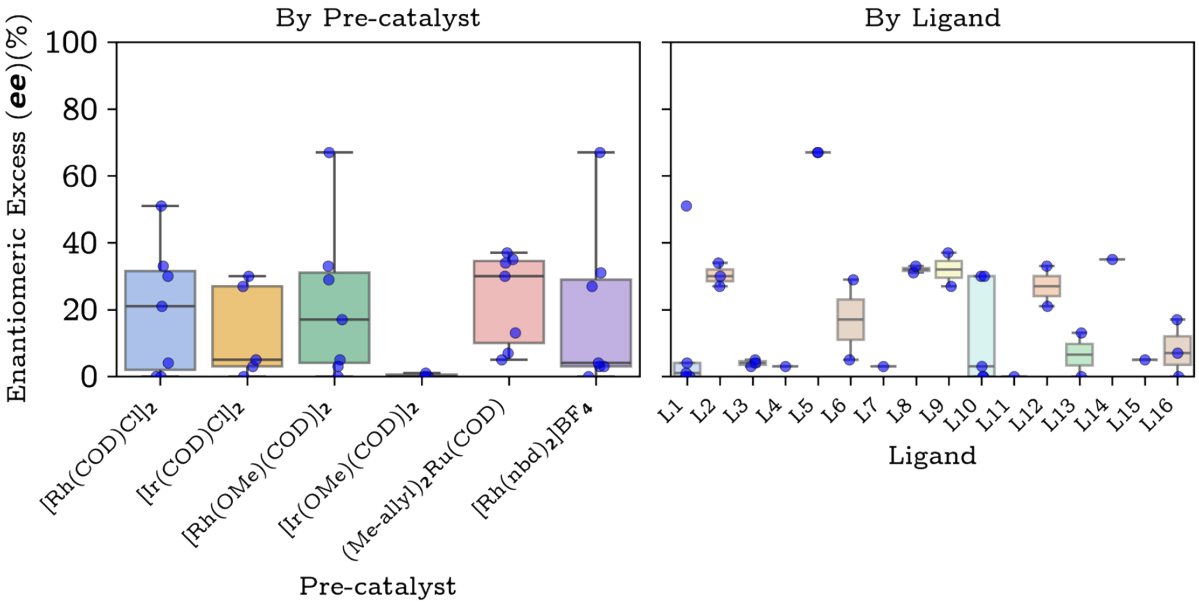


Figure 4.11: Distribution of *ee* across the initial reactions, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.

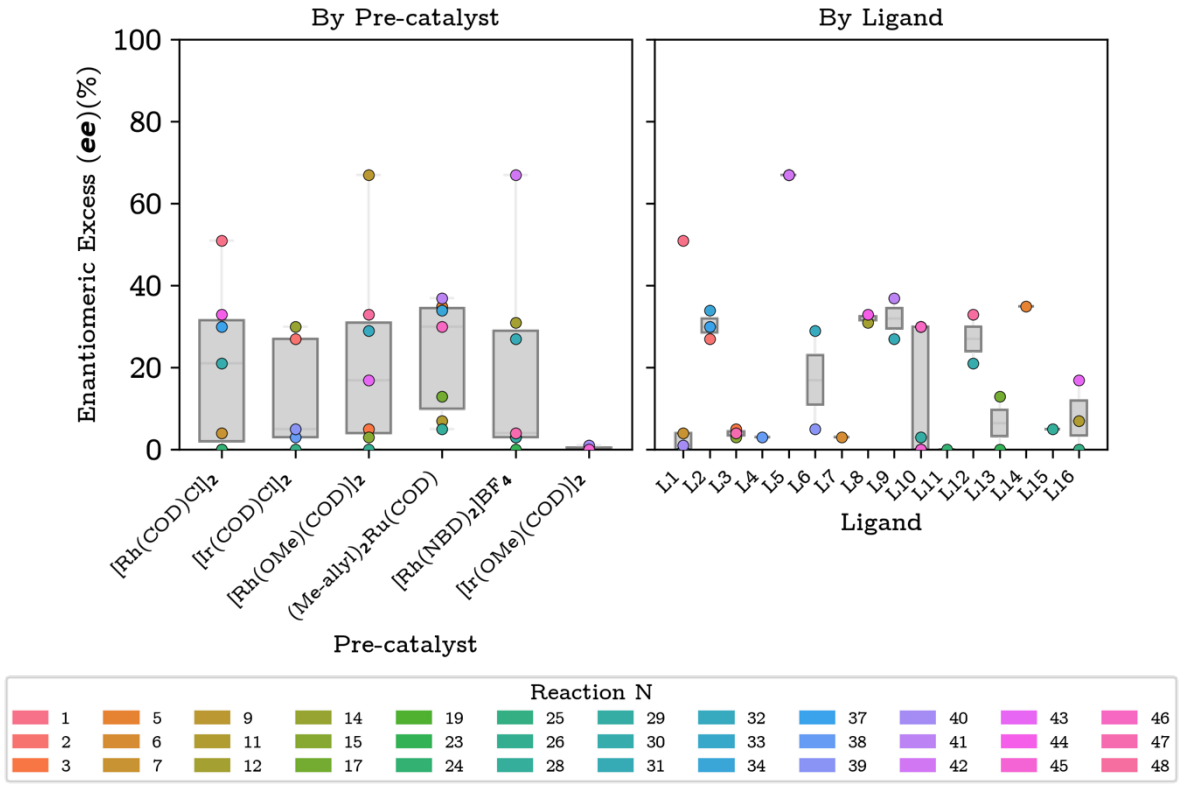
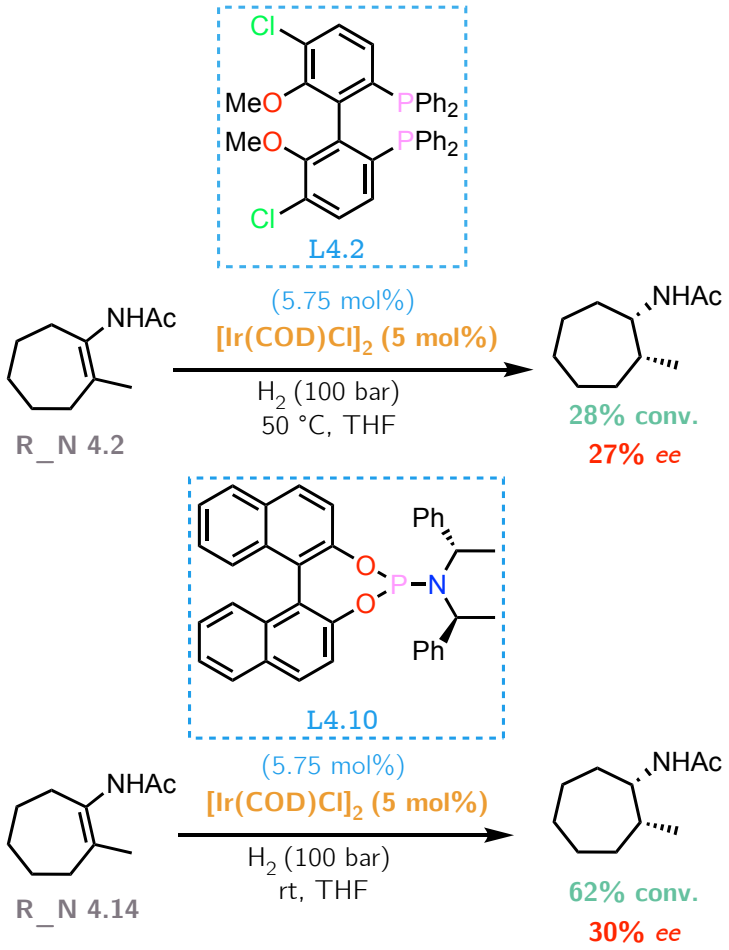
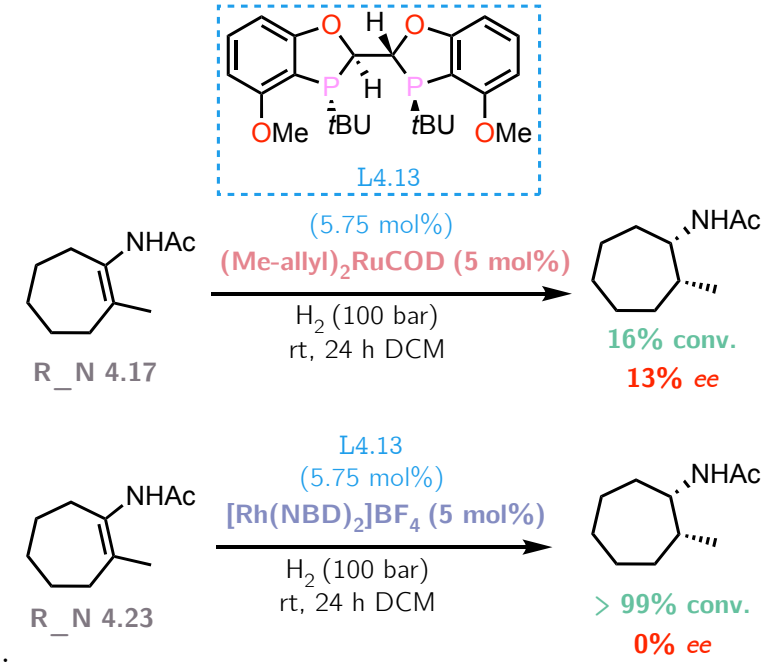


Figure 4.12: Distribution of *ee* across the initial reactions, categorised by pre-catalyst (left) and ligand (right). Each coloured point represents a distinct reaction, with colour indicating reaction identity.



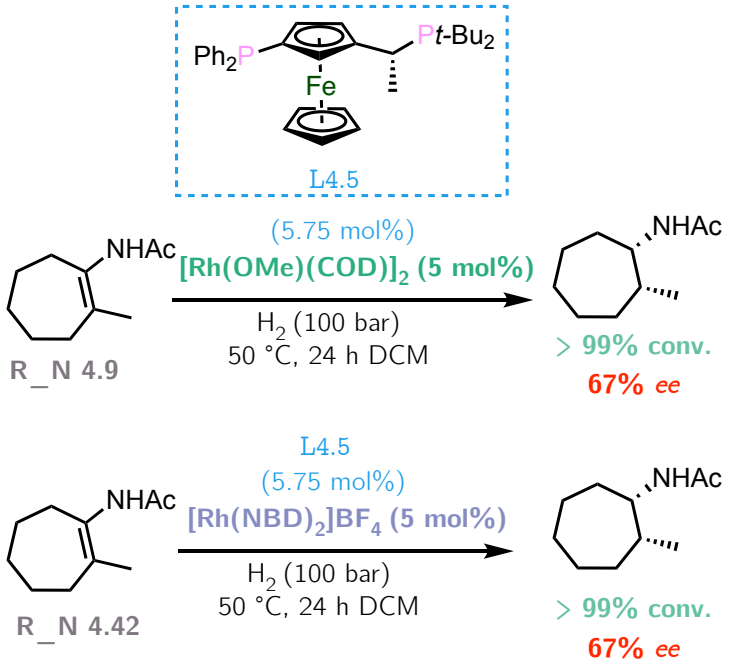
Scheme 4.7: The reactions yielding the highest ee with the pre-catalyst $[Ir(COD)Cl]_2$, using ligands L4.2 and L4.10 in THF at 50 and 25 °C, respectively.

Tang *et al.* reported that a modified version of the commercially available ligand **L4.13** achieved 70% *ee* on *N*-(2-methylcyclohex-1-en-1-yl)acetamide (**Scheme 4.5**). However, for **S4.1**, **L4.13** gave significantly lower enantioselectivity, yielding only 13% *ee* with $(Me\text{-}allyl)_2Ru(COD)$ and 0% *ee* with $[Rh(NBD)_2]BF_4$ in DCM at room temperature, as shown in **Scheme 4.8**.



Scheme 4.8: The reactions conducted with L4.13, and their corresponding conversion and enantioselectivities.

The highest *ee* observed was 67%, achieved in two separate reactions using the pre-catalysts $[Rh(OMe)(COD)]_2$ and $[Rh(NBD)_2]BF_4$ (**Scheme 4.9**). Both reactions employed the same ligand, **L4.5**, and were conducted under identical conditions: 50 °C in DCM with no additive. Despite the use of different pre-catalysts, both systems gave the same *ee*. This 67% *ee* represents a significant improvement over the initial report by Tang *et al.* However, for application in drug discovery, a more enantioselective system was required. More importantly, the diversity of these initial reactions provided a valuable foundation for the BO workflow.



Scheme 4.9: The two initial reactions that achieved the highest ee, both using ligand L4.5 in DCM at 50 °C. The reactions differed only in the pre-catalyst used: $[Rh(OMe)(COD)]_2$ and $[Rh(NBD)_2]BF_4$.

4.3.3 Optimisation One

The results of the 48 initial reactions were input into the BO workflow, and the 24 reactions with the highest EI scores were selected. **Table 4.2** shows the reactions conducted and their corresponding conversion and *ee*. As the top-ranked reaction was set to be carried out at 80 °C all 24 selected reactions were conducted at this temperature, in accordance with the procedure described in **Section 3.2**.

Figure 4.13 illustrates the distribution of *ee* and conversion across the 24 reactions. Nearly all reactions achieved high conversions, with 23 out of 24 reaching full conversion to product, likely a result of the elevated reaction temperature. In contrast, the *ee* values exhibited greater variability. While nine reactions produced an *ee* below 10%, a significant number achieved moderate *ee* values. The average *ee* for this set was 23.0%, representing a notable improvement over the 13.6% observed in the initial reaction set.

Table 4.2: The 24 reactions conducted in Optimisation one. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.49	**L4.14**	$[Ir(COD)Cl]_2$	MeOH		80	100	0
4.50	**L4.16**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	33
4.51	**L4.4**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	22
4.52	**L4.2**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	9
4.53	**L4.3**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	11
4.54	**L4.9**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	31
4.55	**L4.11**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	29
4.56	**L4.7**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	0
4.57	**L4.1**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	11
4.58	**L4.8**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	39
4.59	**L4.12**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	9
4.60	**L4.7**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	0
4.61	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	8
4.62	**L4.4**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	16
4.63	**L4.13**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	13
4.64	**L4.15**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	67
4.65	**L4.11**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	2
4.66	**L4.12**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	46
4.67	**L4.5**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	26	64
4.68	**L4.11**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	0
4.69	**L4.1**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	76
4.70	**L4.11**	$[Rh(COD)Cl]_2$	MeOH		80	100	39
4.71	**L4.8**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	4
4.72	**L4.6**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	31

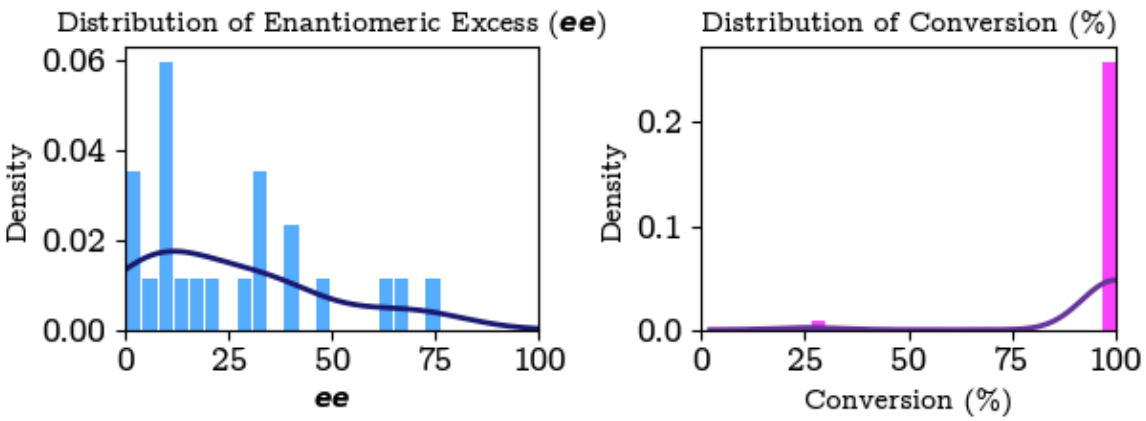


Figure 4.13: The distribution of *ee* and conversion across the set of reactions in Optimisation one.

The top 11 ranked reactions, as identified by the BO workflow, employed the pre-catalyst $[Rh(NBD)_2]BF_4$ in MeOH. Among these, the highest *ee* achieved was 39%, observed with ligand **L4.8**. **Figure 4.14** presents the distribution of *ee* values categorised by both pre-catalyst and ligand. Although the BO algorithm predominantly selected $[Rh(NBD)_2]BF_4$ as the pre-catalyst, a broader diversity of ligands was explored, with 15 out of the 16 available chiral ligands represented. Following $[Rh(NBD)_2]BF_4$, eight reactions were carried out using $(Me\text{-}allyl)_2Ru(COD)$, with a wide range of *ee* values observed across both pre-catalysts and ligand combinations.

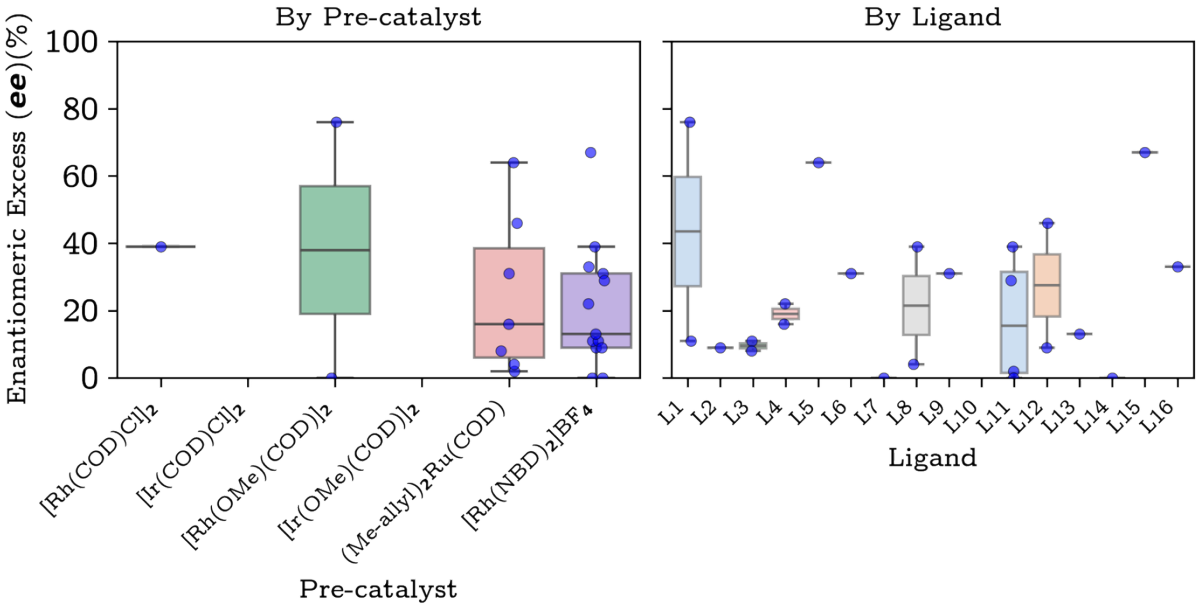
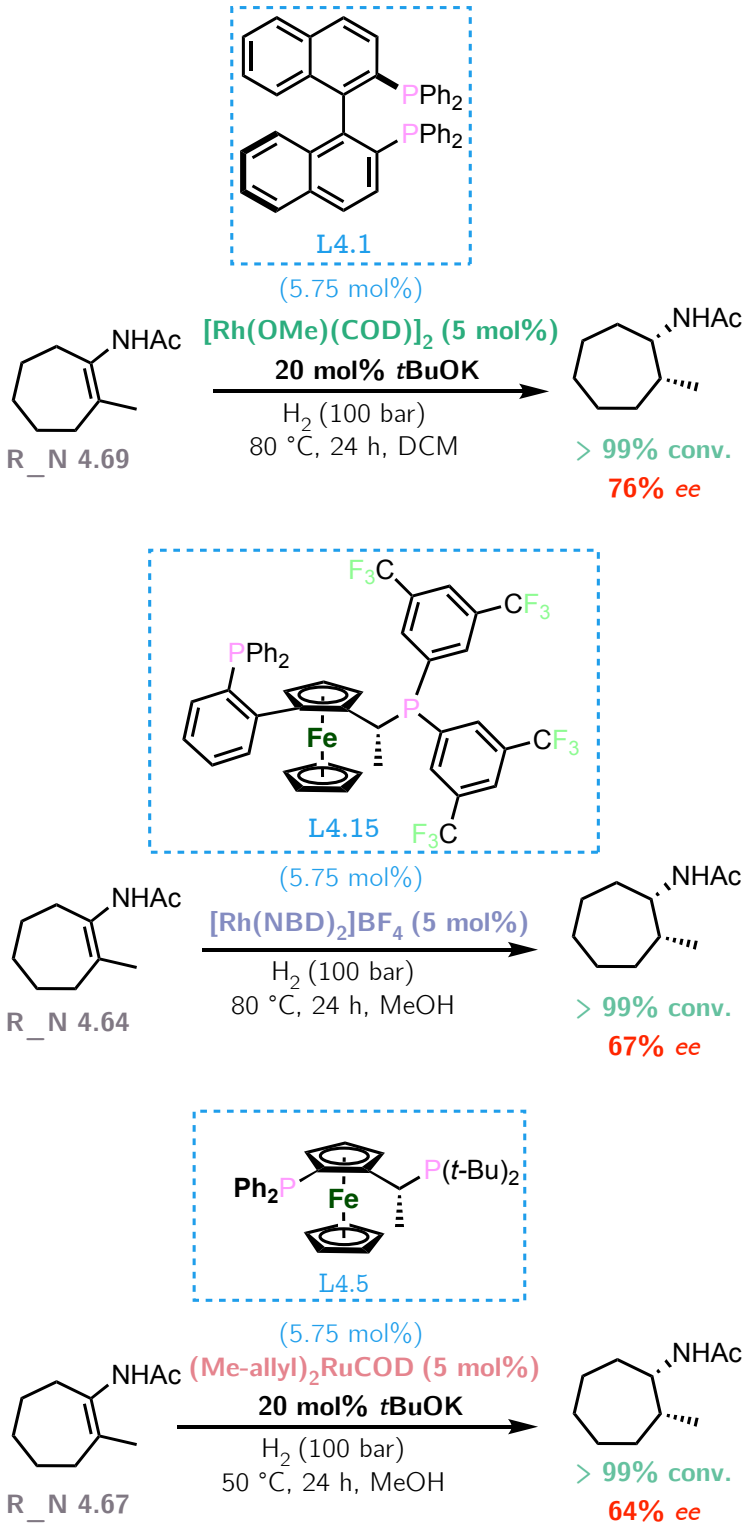


Figure 4.14: Distribution of *ee* across the reactions in Optimisation one, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.

The reactions that produced the highest *ee* showed variation in chiral ligand, pre-catalyst, and solvent, as shown in **Scheme 4.10**. Notably, ligand **L4.5** once again delivered a high *ee* of 64% under a different set of conditions compared to those in **Scheme 4.9**. In this case, the pre-catalyst $(Me\text{-}allyl)_2Ru(COD)$ was used in MeOH with 20 mol% *t*BuOK. This *ee* value is only marginally lower than the 67% previously observed, highlighting the consistent performance of **L4.5** under varied conditions. Interestingly, although **L4.5** showed the best performance in the initial reaction set, it was only selected once by the BO workflow, and that reaction ranked 19th in terms of

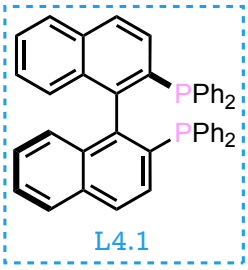
EI score. This, combined with its slightly lower *ee* in the BO-guided reactions, suggests that a reaction resulting in a higher *ee* value for this ligand may not exist in the pre-defined set of 1,782 reactions.



Scheme 4.10: The three reactions in Optimisation one that achieved the highest *ee*.

Another high-performing reaction involved ligand **L4.15**, another ferrocene-based ligand, in combination with $[Rh(NBD)_2]BF_4$ in MeOH, resulting in an *ee* of 67%, matching the highest value observed in the initial reaction set. Notably, one reaction surpassed this benchmark, achieving an *ee* of 76% using (*R*)-Binap (**L4.1**). This reaction employed $[Rh(OMe)(COD)]_2$ as the pre-catalyst with *t*BuOK and was one of only two reactions conducted in DCM instead of MeOH. **Table 4.3** displays the details from the initial set of reactions using **L4.1**, all of which produced significantly lower *ee* values and, in most cases, significantly lower conversions. In this instance, the BO workflow effectively identified an improved set of reaction conditions, most notably the switch to $[Rh(OMe)(COD)]_2$ as a pre-catalyst, resulting in a substantial enhancement in reaction performance for this ligand and substrate combination.

Table 4.3: The highest *ee* reaction and the reactions also using L4.1 from the initial training set.



L4.1

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.69	**L4.1**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	76
4.1	**L4.1**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	62	51
4.7	**L4.1**	$[Rh(COD)Cl]_2$	DCM		50	100	4
4.19	**L4.1**	$[Rh(COD)Cl]_2$	DCM		rt	52	0
4.33	**L4.1**	$[Ir(OMe)(COD)]_2$	MeOH		rt	18	0
4,40	**L4.1**	$[Ir(OMe)(COD)]_2$	THF		80	6	1

4.3.4 Optimisation Two

Following the addition of 24 new data points to the model training set, the next 24 reactions, selected based on the highest EI scores, are listed in **Table 4.4.**

Table 4.4: The 24 reactions conducted in Optimisation two. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.73	**L4.14**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	78
4.74	**L4.16**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	11	4
4.75	**L4.14**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	1
4.76	**L4.10**	$[Rh(OMe)(COD)]_2$	DCM	tBuOK	80	100	3
4.77	**L4.9**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	29
4.78	**L4.12**	$[Rh(COD)Cl]_2$	MeOH		80	30	12
4.79	**L4.1**	$[Rh(COD)Cl]_2$	MeOH		80	100	0
4.80	**L4.15**	$[Rh(COD)Cl]_2$	MeOH		80	100	55
4.81	**L4.5**	$[Rh(COD)Cl]_2$	MeOH		80	100	65
4.82	**L4.2**	$[Rh(COD)Cl]_2$	MeOH		80	100	19
4.83	**L4.14**	$[Rh(COD)Cl]_2$	MeOH		80	100	53
4.84	**L4.16**	$[Ir(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	4
4.85	**L4.6**	$[Rh(COD)Cl]_2$	MeOH		80	100	29
4.86	**L4.10**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	9
4.87	**L4.9**	$[Rh(COD)Cl]_2$	MeOH		80	100	17
4.88	**L4.6**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	100	0
4.89	**L4.4**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	100	11
4.90	**L4.1**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		80	100	49
4.91	**L4.8**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	100	1
4.92	**L4.5**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	100	41
4.93	**L4.13**	$[Rh(COD)Cl]_2$	THF		80	100	21
4.94	**L4.4**	$[Rh(COD)Cl]_2$	MeOH		80	100	7
4.95	**L4.9**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		80	7	0
4.96	**L4.7**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	100	33

As in the previous optimisation round, all reactions were conducted at 80 °C, resulting in consistently high conversions, 21 out of 24 reactions achieved full conversion to product. Similar to Optimisation one, a wider variation in *ee* was observed, with a slightly lower mean *ee* of 22.6% (**Figure 4.15**).

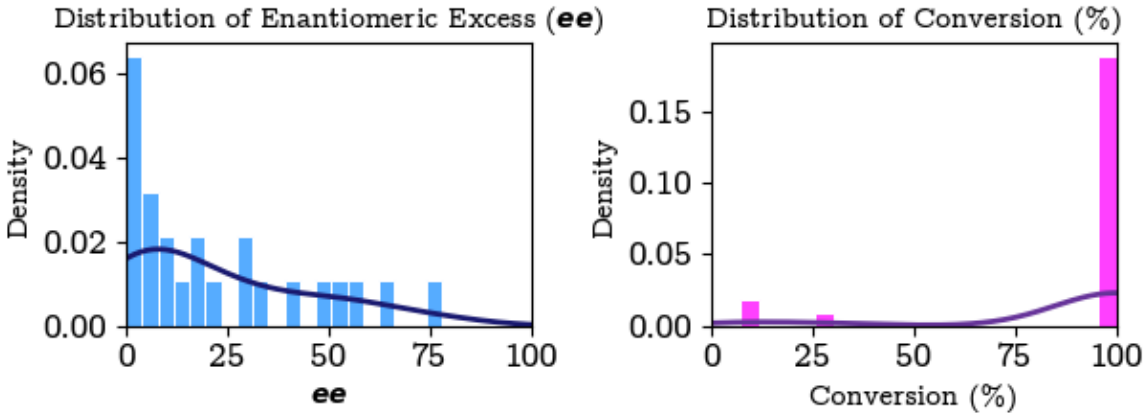


Figure 4.15: The distribution of *ee* and conversion across the set of reactions in Optimisation two.

Once again, a broad range of ligands was selected, with 14 out of the 16 chiral ligands used (**Figure 4.16**). In contrast, there was less variation in pre-catalysts: 15 of the 24 reactions were conducted using $[Rh(COD)Cl]_2$, a pre-catalyst that was less explored during Optimisation one. This suggests that the AF is prioritising exploration across diverse reaction parameters rather than refining conditions around the previously best-performing ligand, **L4.1**. Notably, the two reactions involving **L4.1** in this round (reactions 4.79 and 4.90) performed worse than the optimal pre-catalyst identified earlier, with *ee* values of 0% and 49%, when using $[Rh(COD)Cl]_2$ and $(Me\text{-}allyl)_2Ru(COD)$, respectively. These results contrast with the 76% *ee* previously achieved using $[Rh(OMe)(COD)]_2$ under the same reaction conditions of 80 °C in MeOH.

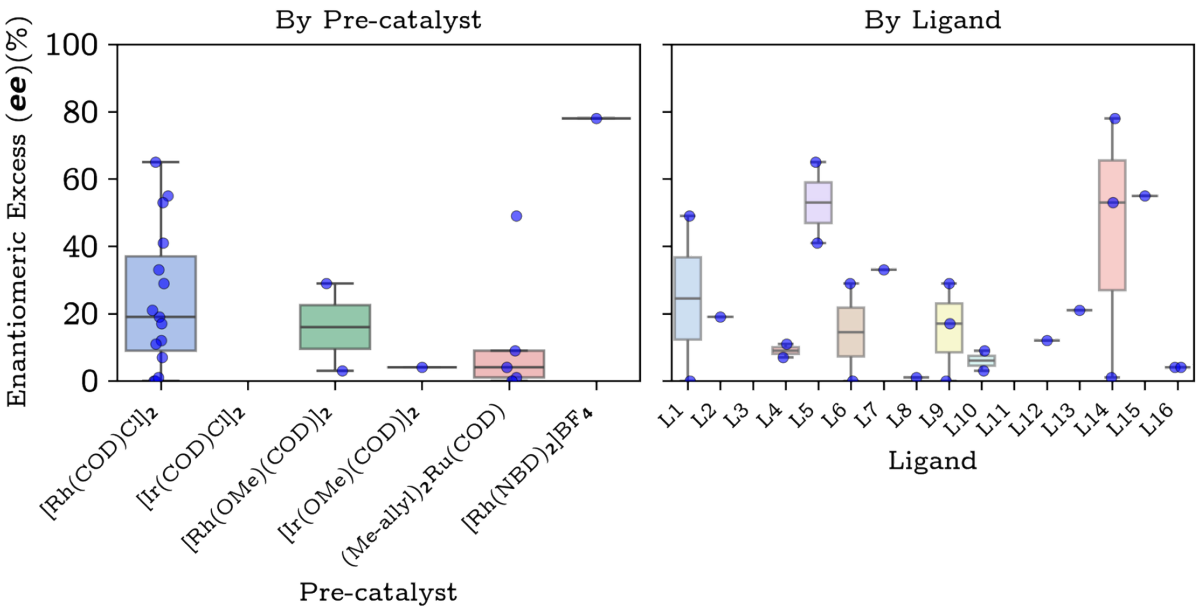
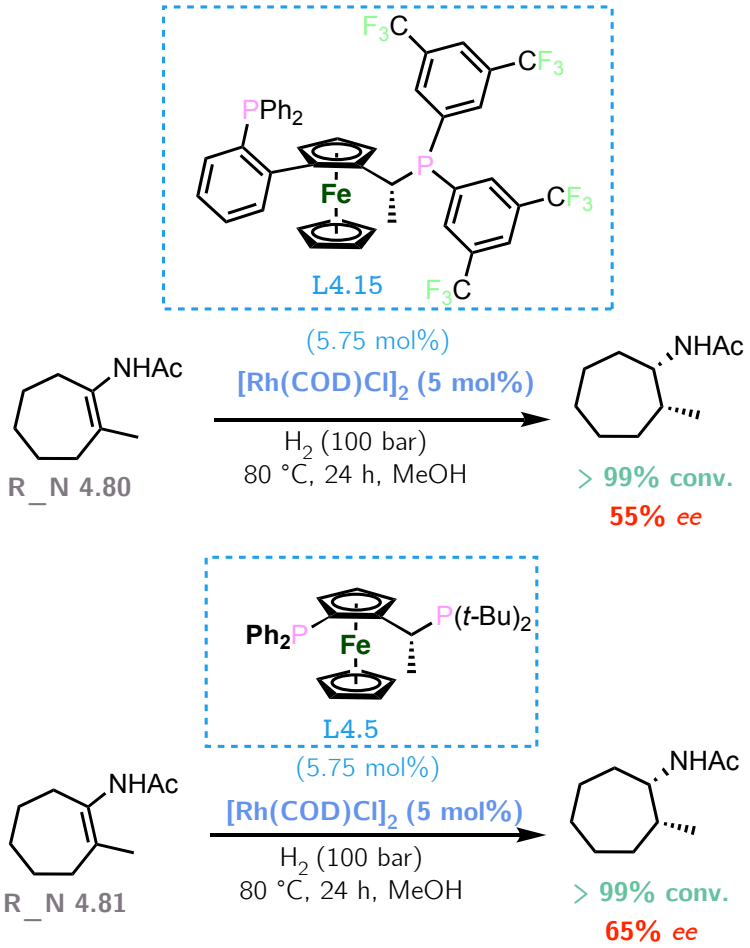


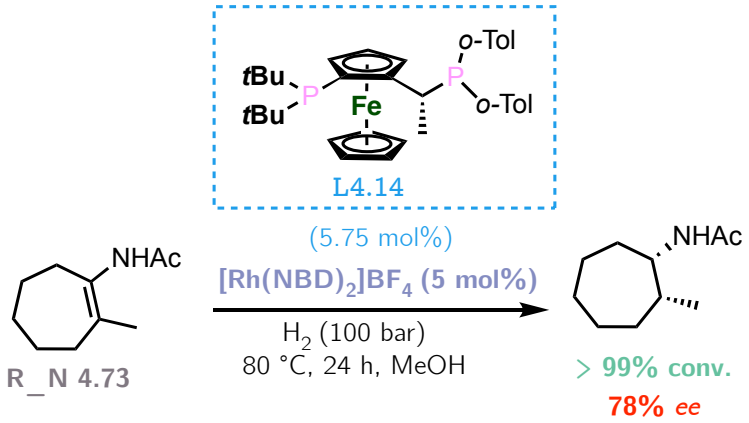
Figure 4.16: Distribution of *ee* across the reactions in Optimisation two, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.

Once again, **L4.5**, demonstrated consistent performance, achieving 65% *ee* with $[Rh(COD)Cl]_2$ in MeOH at 80 °C (**Scheme 4.11**). However, this result did not surpass the previously observed 67% *ee*, obtained with $[Rh(OMe)COD]_2$ or $[Rh(NBD)2]BF_4$ in DCM at 50 °C (**Scheme 4.9**). A similar result was seen for reaction 4.80, where **L4.15** generated a lower 55% *ee* with $[Rh(COD)Cl]_2$ in MeOH at 80 °C, than the 67% observed with $[Rh(NBD)_2]BF_4$ under otherwise identical conditions (**Scheme 4.10**, reaction 4.64).



Scheme 4.11: Reactions 4.80 and 4.81 from Optimisation two.

One reaction from Optimisation two identified a new set of conditions that resulted in a higher *ee* than previously observed (**Scheme 4.12**). This reaction, employing **L4.14** with $[Rh(NBD)_2]BF_4$ in MeOH at 80 °C, outperformed all earlier conditions tested in both the initial reaction set and Optimisation one, demonstrating the ability of the workflow to discover improved reaction parameters.



Scheme 4.12: The reaction resulting in the highest *ee* in Optimisation two.

4.3.5 Optimisation Three

While in Optimisations one and two, reactions yielding a higher *ee* were found each time, Optimisation three was not as successful. The reactions conducted and their results are displayed in **Table 4.5**, and the distribution of the conversion and *ee* is shown in **Figure 4.17**. A narrower distribution of both *ee* and conversion was observed, with five reactions showing no conversion to product, eight resulting in racemic mixtures, and a further five showed less than 3% *ee*. As a result, the mean *ee* was significantly reduced to 4.4%.

Table 4.5: The 24 reactions conducted in Optimisation three. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.97	**L4.16**	$[Rh(COD)Cl]_2$	MeOH		80	100	21
4.98	**L4.6**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	0
4.99	**L4.16**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	100	0
4.00	**L4.10**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	90	0
4.101	**L4.16**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		80	0	-
4.102	**L4.3**	$[Rh(NBD)_2]BF_4$	DCM	NEt_3	80	100	2
4.103	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH		80	100	10
4.104	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	THF	NEt_3	80	100	2
4.105	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	DCM	NEt_3	80	100	0
4.106	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	NEt_3	80	100	17
4.107	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	THF	*t*BuOK	80	100	6
4.108	**L4.3**	$(Me\text{-}allyl)_2Ru(COD)$	DCM	*t*BuOK	80	100	0
4.109	**L4.3**	$[Ir(OMe)(COD)]_2$	DCM		80	53	2

4.110	**L4.3**	$[Ir(OMe)(COD)]_2$	MeOH		80	0	-
4.111	**L4.3**	$[Ir(OMe)(COD)]_2$	THF	NEt_3	80	0	-
4.112	**L4.3**	$[Ir(OMe)(COD)]_2$	DCM	NEt_3	80	44	2
4.113	**L4.3**	$[Ir(OMe)(COD)]_2$	MeOH	NEt_3	80	6	0
4.114	**L4.3**	$[Ir(OMe)(COD)]_2$	THF	*t*BuOK	80	0	-
4.115	**L4.9**	$[Ir(COD)Cl]_2$	DCM		80	100	2
4.116	**L4.9**	$[Ir(COD)Cl]_2$	MeOH		80	100	19
4.117	**L4.9**	$[Rh(COD)Cl]_2$	MeOH	*t*BuOK	80	100	0
4.118	**L4.3**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	80	79	1
4.119	**L4.3**	$[Rh(COD)Cl]_2$	MeOH	*t*BuOK	80	100	0
4.120	**L4.8**	$[Rh(NBD)_2]BF_4$	THF		80	0	-

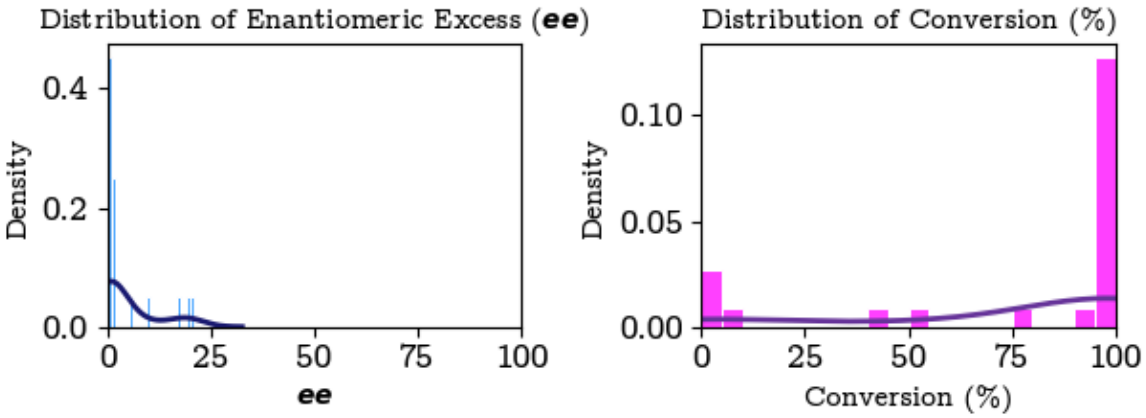


Figure 4.17: The distribution of *ee* and conversion across the set of reactions in Optimisation three.

Figure 4.18 illustrates the distribution of *ee* values grouped by both pre-catalyst and ligand. Compared to Optimisations one and two, Optimisation three exhibited less diversity in the ligands selected, while all available pre-catalysts were represented. 15 reactions employed **L4.3**, all of which showed poor performance, with the highest observed *ee* reaching only 17%.

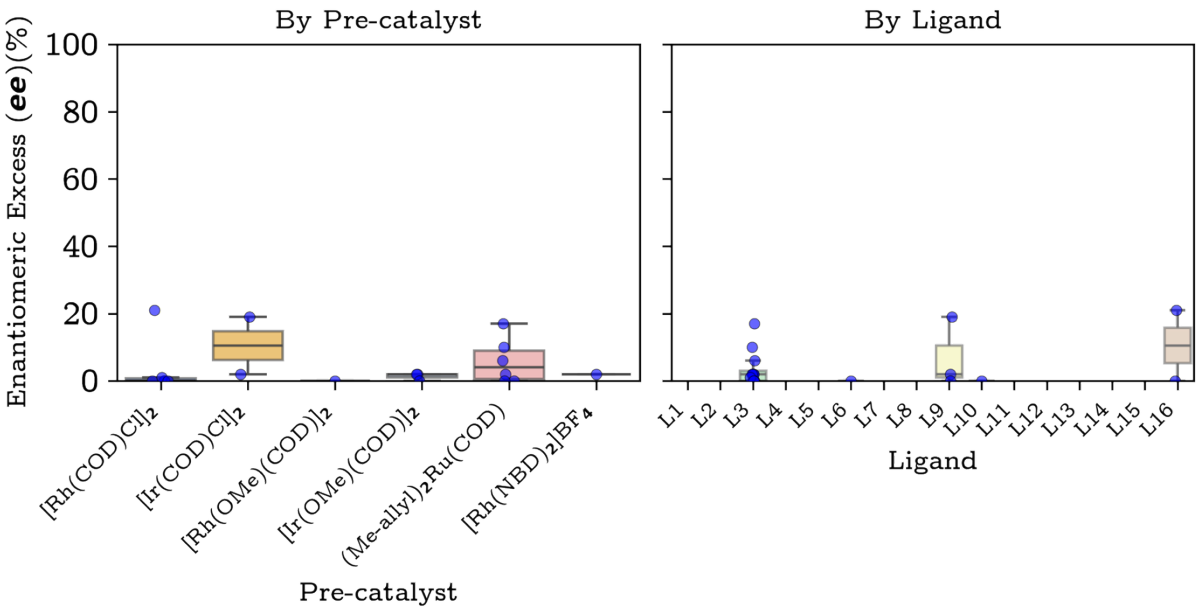
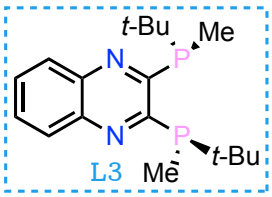


Figure 4.18: Distribution of *ee* across the reactions in Optimisation two, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.

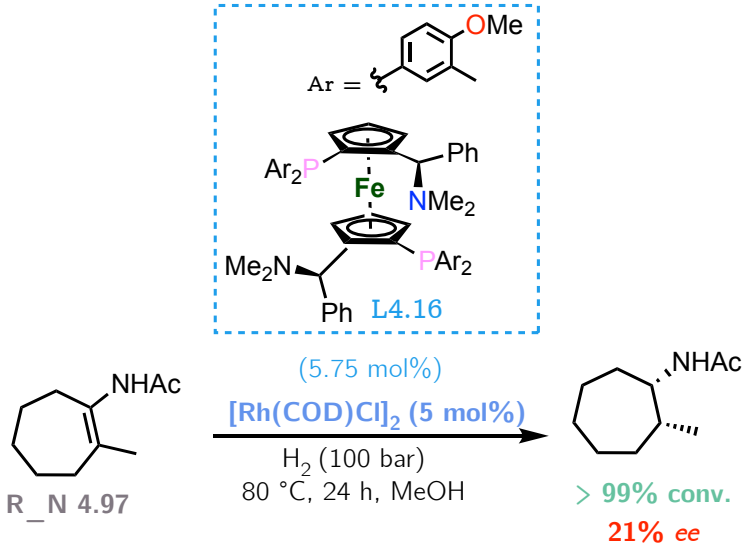
Table 4.6 presents the reactions that employed **L4.3** in the initial reaction set, along with its selection in Optimisation one (reaction 4.53). Among these, two reactions showed no conversion to product, and reaction 4.53 achieved only a modest *ee* of 11%. However, reaction 4.3 yielded a more promising *ee* of 51%. Since reactions with no measurable *ee* are excluded from the training set and instead removed from the reaction space, **L4.3** may have appeared as an underexplored region. This could explain why it received a higher EI score than its previous performance would indicate.

Table 4.6: The reactions also using L4.3 from the initial training set (Reactions 4.3, 4.15, 4.21, 4.47) and Optimisation one (reaction 4.53).



R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.3	**L4.3**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	50	62	51
4.15	**L4.3**	$[Rh(OMe)(COD)]_2$	DCM	NEt_3	rt	0	-
4.21	**L4.3**	$[Rh(OMe)(COD)]_2$	DCM		rt	0	-
4.47	**L4.3**	$[Rh(NBD)_2]BF_4$	DCM		80	99	4
4.53	**L4.3**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	11

The reaction with the highest EI score in this optimisation round also delivered the highest *ee* observed, at 21% (**Scheme 4.13**). This reaction employed **L4.16** with $[Rh(COD)Cl]_2$ in MeOH at 80 °C. As it ranked highest in terms of EI and used 80 °C as the reaction temperature, this temperature was adopted for all subsequent reactions in this round. However, it is possible that the high temperature contributed to the overall poorer performance observed in Optimisation three, suggesting that this parameter may be over explored.



Scheme 4.13: The reaction resulting in the highest *ee* in Optimisation three.

The poorer performance observed in Optimisation three highlighted a potential issue with the batch size used in this study. Since the model is only able to 'learn' between optimisation rounds, employing a relatively large batch size of 24 may have resulted in more reactions being performed per iteration than necessary. This could lead to inefficient exploration of the reaction space. For example, the repeated selection of **L4.3**, despite its poor performance, illustrates how a large batch size may delay the model's ability to effectively redirect its focus based on new data. For this reason, the batch size was reduced from 24 to 12 in subsequent optimisation rounds.

4.3.6 Optimisation Four

The 12 reactions selected for Optimisation four, and their corresponding experimental results, are shown in **Table 4.7**. The reaction temperature was reduced to 50 °C, based on the conditions of reaction 4.121, the reaction with the highest EI score in this round. The distribution of *ee* and conversion for this set is shown in **Figure 4.19**. Compared to the previous optimisation rounds, overall conversion was lower, likely as a result of the decreased temperature, with only four of the 12 reactions reaching full conversion to product.

Although four reactions resulted in racemic mixtures, several reactions exhibited mid-range *ee* values, leading to a higher mean *ee* of 29.0%, the highest observed across all optimisation rounds to date.

Table 4.7: The 12 reactions conducted in Optimisation four. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.121	**L4.16**	$[Rh(OMe)(COD)]_2$	DCM		50	100	9
4.122	**L4.14**	$[Rh(OMe)(COD)]_2$	DCM		50	65	51
4.123	**L4.4**	$[Rh(OMe)(COD)]_2$	DCM		50	100	0
4.124	**L4.11**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	0	-
4.125	**L4.2**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	43	63
4.126	**L4.12**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	61	69
4.127	**L4.10**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	100	0
4.128	**L4.7**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	32	1
4.129	**L4.4**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	97	0

4.130	**L4.3**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	18	54
4.131	**L4.2**	$[Rh(OMe)(COD)]_2$	DCM		50	100	51
4.132	**L4.5**	$[Rh(COD)Cl]_2$	DCM	tBuOK	50	49	21

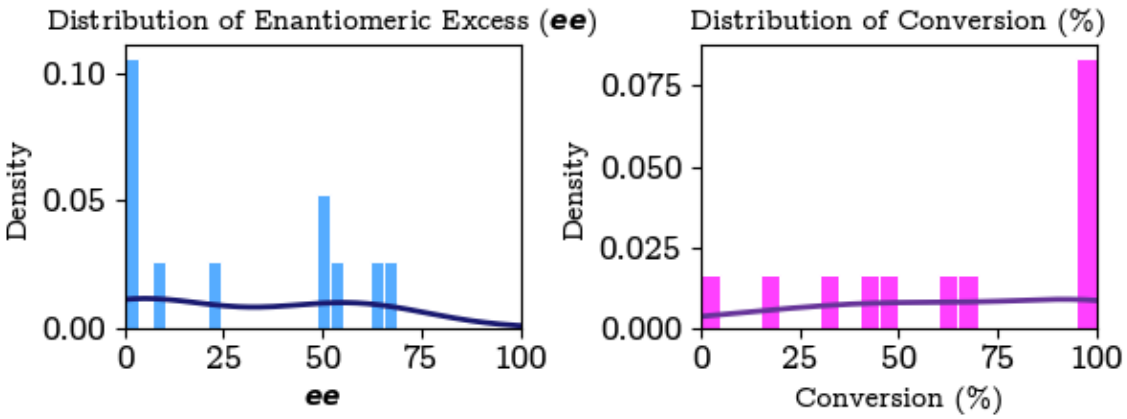


Figure 4.19: The distribution of *ee* and conversion across the set of reactions in Optimisation four.

Figure 4.20 categorises the *ee* values by pre-catalyst and ligand for this optimisation round. There was substantial variation in ligand selection, with only one ligand used more than once among the 12 reactions. In contrast, pre-catalyst selection was more limited, with all reactions employing either $[Rh(COD)Cl]_2$ or $[Rh(OMe)(COD)]_2$. Consistently, all reactions were conducted in DCM, with eight using *t*BuOK as an additive and the remaining four conducted without any additive.

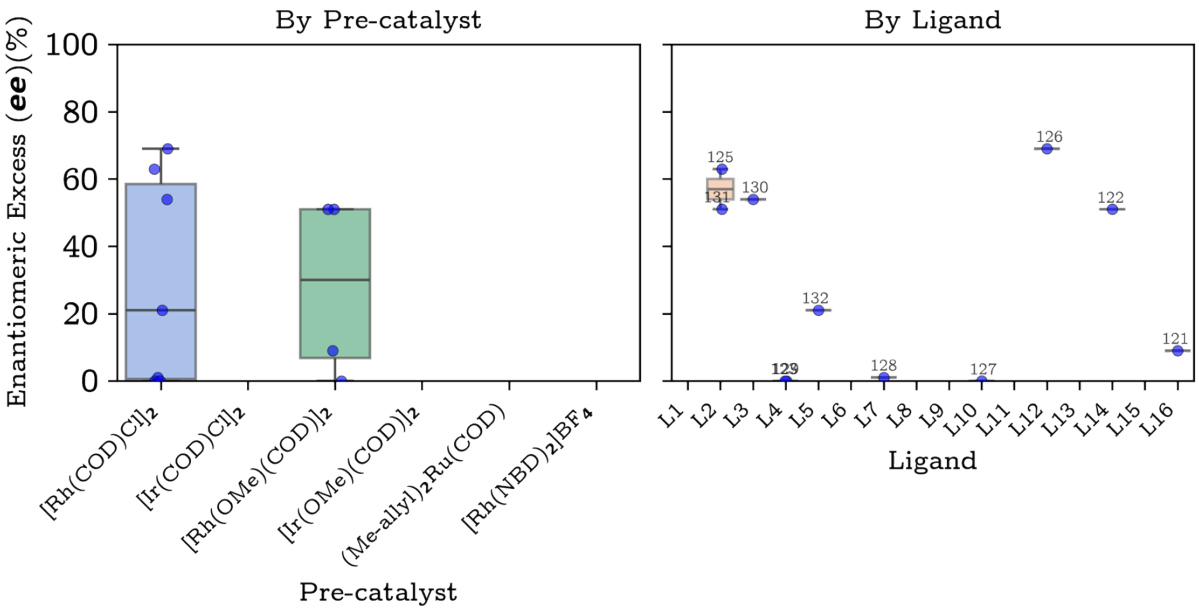
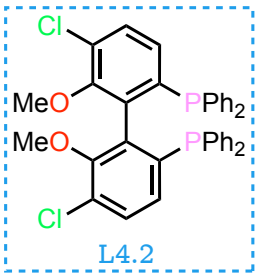


Figure 4.20: Distribution of *ee* across the reactions in Optimisation four, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.

Despite the increased average *ee* observed in this set of reactions, no individual reaction exceeded the previously recorded maximum *ee* of 78%. Nonetheless, mid-range *ee* values were achieved in Reactions 4.125 and 4.126, which employed chiral ligands **L4.2** and **L4.12**, respectively, in combination with $[Rh(COD)Cl]_2$ and *t*BuOK, resulting in *ee* values of 62% and 69%. Tables **4.8** and **4.9** display the reactions from the initial set and previous optimisation rounds involving **L4.2** and **L4.12**, which formed part of the training data for Optimisation four. For both ligands, improved reaction conditions were identified, resulting in higher *ee* values compared to those in the training set. This demonstrates that the BO workflow is capable of identifying reaction conditions that enhance enantioselectivity.

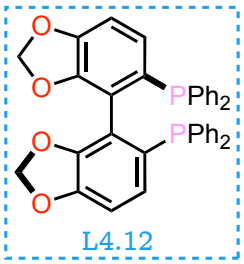
Table 4.8: The reactions also using L4.2 from the initial training set (Reactions 4.2, 4.20, 4.34, 4.37), Optimisation one (reaction 4.52) and Optimisation two (reaction 4.82).



L4.2

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	ee(%)
4.125	**L4.2**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	43	63
4.2	**L4.2**	$[Ir(COD)Cl]_2$	THF		50	28	27
4.20	**L4.2**	$[Ir(COD)Cl]_2$	THF	NEt_3	rt	0	-
4.34	**L4.2**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		rt	100	34
4.37	**L4.2**	$[Rh(COD)Cl]_2$	THF	*t*BuOK	80	67	30
4.52	**L4.2**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	9
4.82	**L4.2**	$[Rh(COD)Cl]_2$	MeOH		80	100	19

Table 4.9: The reactions also using L4.12 from the initial training set (Reactions 4.16, 4.30 and 4.48), Optimisation one (reaction 4.59 and 4.66) and Optimisation two (reaction 4.78).



L4.12

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	ee(%)
4.126	**L4.12**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	61	69
4.16	**L4.12**	$[Ir(OMe)(COD)]_2$	THF		rt	0	-
4.30	**L4.12**	$[Rh(COD)Cl]_2$	MeOH	NEt_3	rt	100	21
4.48	**L4.12**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	99	33
4.59	**L4.12**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	9
4.66	**L4.12**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	46
4.78	**L4.12**	$[Rh(COD)Cl]_2$	MeOH		80	30	12

4.3.7 Optimisation Five

The reactions conducted in Optimisation five and their corresponding results are presented in **Table 4.10**, with the distribution of *ee* and conversion shown in **Figure 4.21**. As in Optimisation four, all reactions were carried out at 50 °C. Compared to the previous round, higher conversions were observed, with eight out of the 12 reactions reaching full conversion to product. Notably, the overall performance in this round was substantially improved, with a mean *ee* of 44.7%, and only two reactions exhibiting *ee* values below 20%.

Table 4.10: The 12 reactions conducted in Optimisation five. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.133	**L4.6**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	100	10
4.134	**L4.14**	$[Rh(COD)Cl]_2$	DCM	*t*BuOK	50	70	33
4.135	**L4.12**	$[Rh(OMe)(COD)]_2$	DCM		50	63	67
4.136	**L4.6**	$[Rh(OMe)(COD)]_2$	DCM		50	100	23
4.137	**L4.1**	$[Rh(OMe)(COD)]_2$	DCM		50	52	59
4.138	**L4.2**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	47
4.139	**L4.9**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	61
4.140	**L4.11**	$[Rh(NBD)_2]BF_4$	THF		50	100	55
4.141	**L4.10**	$[Rh(OMe)(COD)]_2$	DCM		50	100	19
4.142	**L4.14**	$[Rh(NBD)_2]BF_4$	THF		50	100	93
4.143	**L4.2**	$[Rh(NBD)_2]BF_4$	THF		50	100	24
4.144	**L4.3**	$[Rh(COD)Cl]_2$	DCM		50	0	-

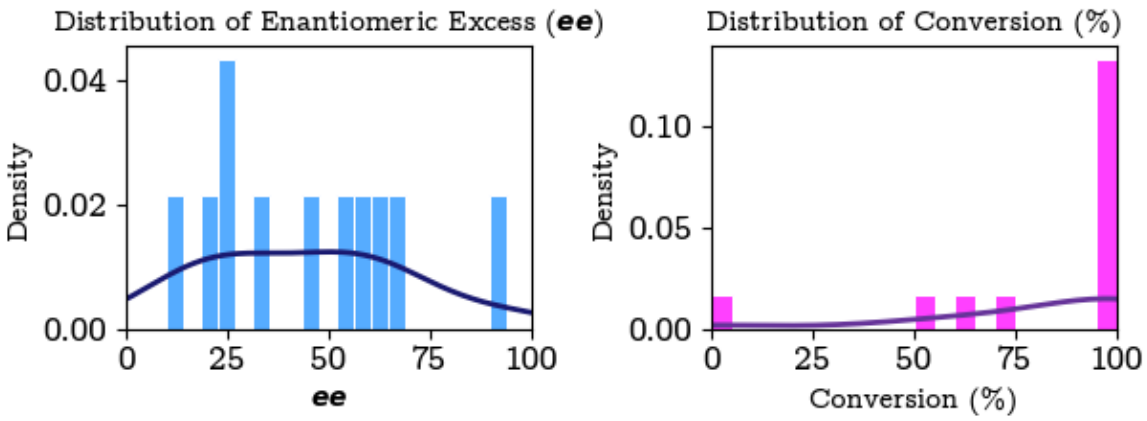


Figure 4.21: The distribution of *ee* and conversion across the set of reactions in Optimisation five.

The distribution of *ee* values, categorised by both pre-catalyst and ligand, is shown in **Figure 4.22**. A high degree of variation was observed in the selection of both pre-catalysts and ligands, with nine different chiral ligands and four pre-catalysts employed across the 12 reactions. In contrast, and consistent with previous optimisation rounds, less variation was seen in the choice of solvent; nine reactions were conducted in DCM, while the remaining three used THF.

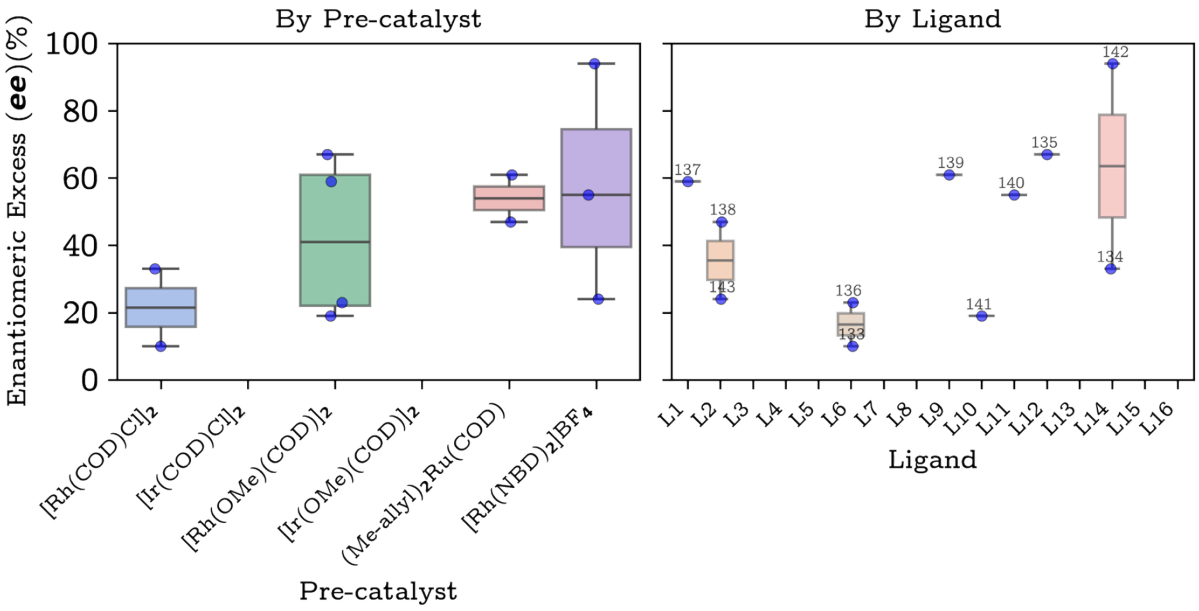
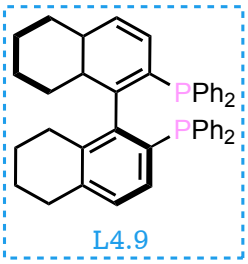


Figure 4.22: Distribution of *ee* across the reactions in Optimisation five, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.

L4.12 continued to show consistent performance, seen in reaction 4.135, which achieved an *ee* of 67%, closely matching the result from reaction 4.126 in Optimisation four. This reaction also employed $[Rh(OMe)(COD)]_2$ in DCM, but notably, it was conducted without the use of an additive. In addition, the optimisation approach successfully identified improved reaction conditions for a different chiral ligand, **L4.9**, in reaction 4.139. **Table 4.11** presents the previous reactions involving **L4.9** that were included in the training set for comparison. The result from reaction 4.139 demonstrated a substantial improvement, with an *ee* significantly higher than those previously observed, with the best prior result for **L4.9** being 37%.

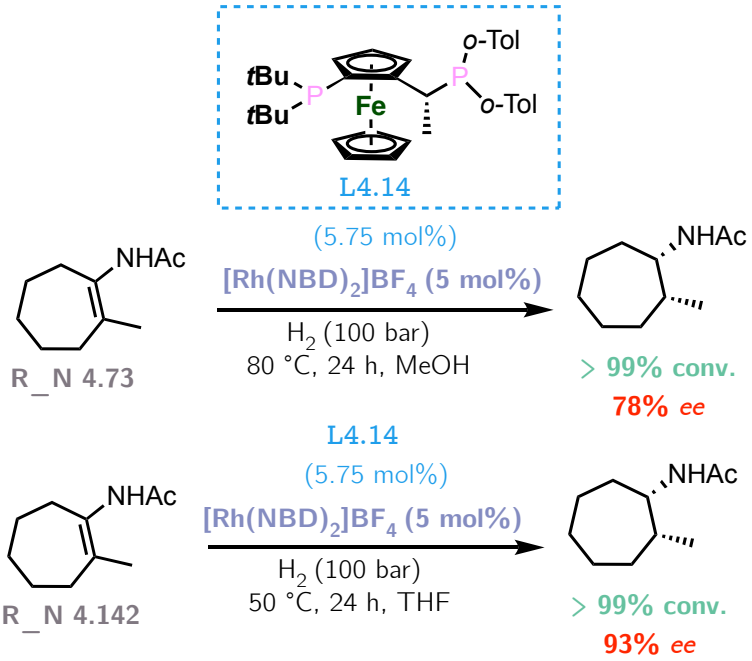
Table 4.11: The reactions also using L4.9 from the initial training set (Reactions 4.8, 4.31 and 4.41), Optimisation one (reaction 4.54), Optimisation two (reaction 4.77, 4.87 and 4.95) and Optimisation three (reactions 4.115, 4.116 and 4.117).



L4.9

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.139	**L4.9**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	61
4.8	**L4.9**	$[Ir(COD)Cl]_2$	THF		50	0	-
4.31	**L4.9**	$[Rh(NBD)_2]BF_4$	MeOH		rt	100	27
4.41	**L4.9**	$(Me\text{-}allyl)_2Ru(COD)$	MeOH	*t*BuOK	80	100	37
4.54	**L4.9**	$[Rh(NBD)_2]BF_4$	MeOH		80	100	31
4.77	**L4.9**	$[Rh(OMe)(COD)]_2$	DCM	*t*BuOK	80	100	29
4.87	**L4.9**	$[Rh(COD)Cl]_2$	MeOH		80	100	17
4.95	**L4.9**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		80	7	0
4.115	**L4.9**	$[Ir(COD)Cl]_2$	DCM		80	100	2
4.116	**L4.9**	$[Ir(COD)Cl]_2$	MeOH		80	100	19
4.117	**L4.9**	$[Rh(COD)Cl]_2$	MeOH	*t*BuOK	80	100	0

The most notable result from this optimisation campaign was observed in reaction 4.142, employing **L4.14**, which achieved a high *ee* of 93% (**Scheme 4.14**). This represents a significant improvement over the 78% *ee* previously obtained with this ligand in Optimisation two, the highest *ee* recorded prior to this point. The reduction in temperature from 80 °C to 50 °C, combined with a change in solvent from MeOH to THF, led to a substantial increase in enantioselectivity, resulting in a notably high *ee* value.



Scheme 4.14: The reactions yielding the highest *ee* values in this optimisation campaign. Reaction 4.73 from Optimisation two and reaction 4.73 from Optimisation five.

4.3.8 Summary of Bayesian Optimisation Performance

As the objective of employing BO to optimise the AH of **S4.1** was to do so in a more data-efficient manner compared to traditional screening approaches, an experimental budget of 10% of the defined reaction space was applied. This corresponded to a limit of 178 reactions to identifying the most selective reaction conditions within a total space of 1,782 possible reactions. As the most selective conditions may have been identified in reaction 4.142, representing just 8% of the total reaction space, one additional optimisation round was conducted to explore the possibility of discovering an even higher *ee* reaction within this iteration.

The reactions conducted in this iteration, along with their corresponding experimental results, are presented in **Table 4.12**, with the distribution of *ee* and conversion shown in **Figure 4.23**. All reactions were carried out at 50 °C, in either DCM or THF. While the *ee* values were generally lower than those achieved in Optimisation five, with four reactions resulting in racemic mixtures and the mean *ee* reducing to 24.8%, conversion remained high, with nine out of the 12 reactions reaching full conversion to product.

Table 4.12: The 12 reactions conducted in Optimisation six. All reactions were carried out under conditions: 100 bar H_2, 24 h, 5 mol% pre-catalyst, 5.75 mol% ligand and 20 mol% additive.

R_N	Ligand	Pre-cat Name	Solvent	Additive	Temp. (°C)	Conv.(%)	*ee*(%)
4.145	**L4.5**	$[Rh(NBD)_2]BF_4$	THF		50	100	73
4.146	**L4.4**	$[Rh(NBD)_2]BF_4$	THF		50	100	8
4.147	**L4.1**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	59
4.148	**L4.11**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	50	2
4.149	**L4.15**	$[Rh(NBD)_2]BF_4$	THF		50	100	59
4.150	**L4.16**	$[Rh(NBD)_2]BF_4$	THF		50	100	27
4.151	**L4.4**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	3
4.152	**L4.7**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	0
4.153	**L4.10**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	0
4.154	**L4.8**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	100	0
4.155	**L4.15**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	13	0
4.156	**L4.13**	$(Me\text{-}allyl)_2Ru(COD)$	DCM		50	16	66

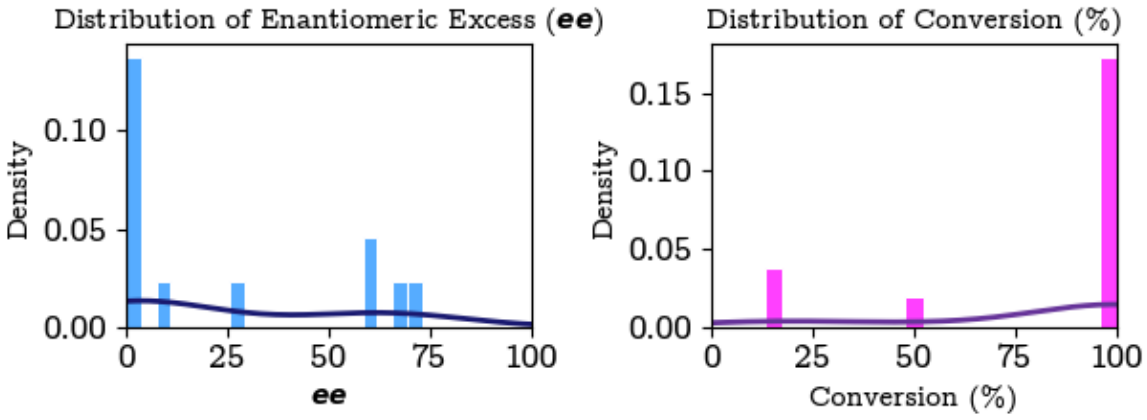


Figure 4.23: The distribution of *ee* and conversion across the set of reactions in Optimisation six.

The distribution of *ee* values categorised by pre-catalyst and ligand for this optimisation round is shown in **Figure 4.24**. Interestingly, although a range of ligands were explored, the best-performing ligand overall, **L4.14**, was not included among the 12 reactions with the highest EI scores. Instead, the highest-ranking reaction (reaction 4.145) employed ligand **L4.5** and achieved the highest *ee* in this final optimisation round (**Scheme 4.19**). This represented an improvement over the previous best result for **L4.5** of 67% *ee* observed in Reactions 4.9 and 4.42 (**Scheme 4.15**). By switching

from switching from DCM to THF, the *ee* increased to 73%. Notably, these are the same reaction conditions used in reaction 4.142, which yielded the highest *ee* overall.

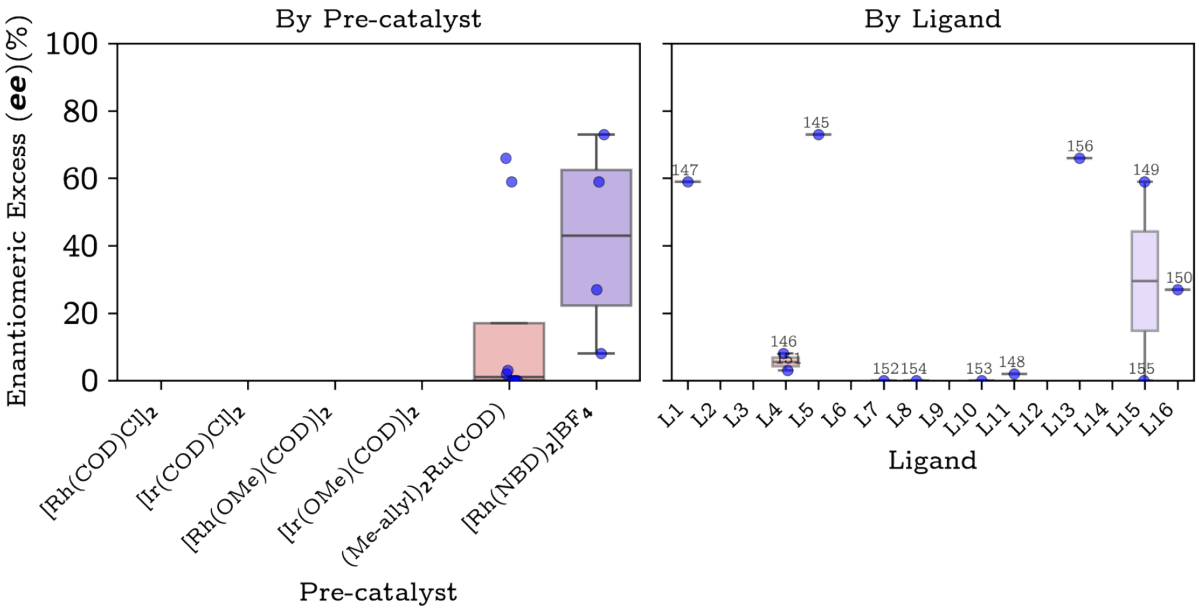
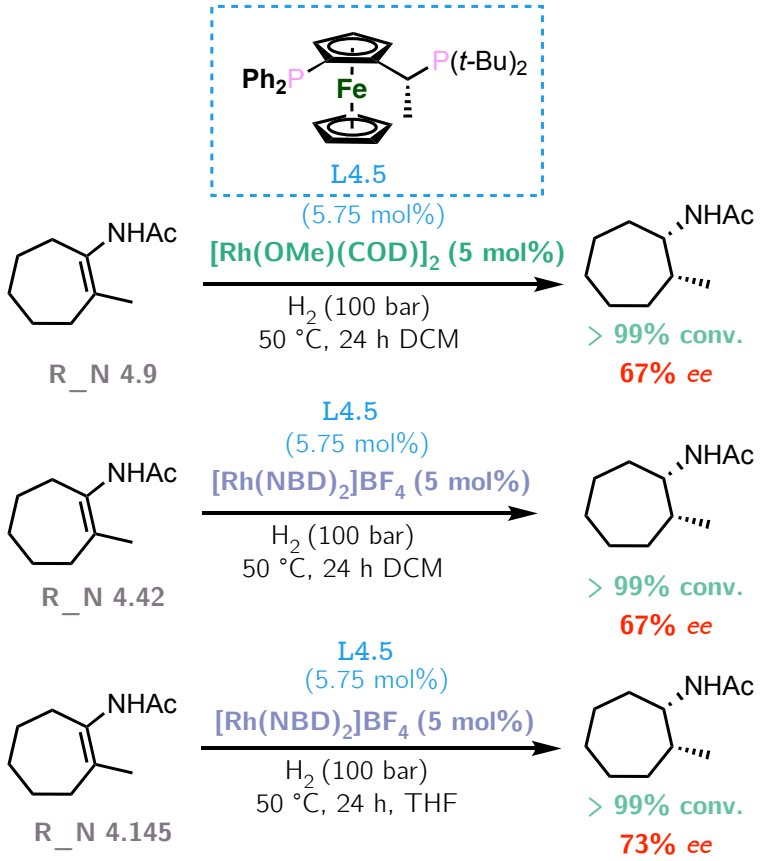


Figure 4.24: Distribution of *ee* across the reactions in Optimisation six, categorised by pre-catalyst (left) and ligand (right). Box plots summarize the *ee* distribution for each category.



Scheme 4.15: Comparison of the reaction that achieved the highest *ee* in Optimisation six with the previously highest-performing reaction using the same ligand.

The results from Optimisation six suggested that the BO workflow was unlikely to identify a reaction more selective than the 93% *ee* achieved in Optimisation five. Given this outcome, and although the experimental budget was not exceeded, the optimisation campaign was concluded without further iterations.

Figure 4.25 shows the distributions of *ee* and conversion for all six optimisation rounds, together with a scatter plot illustrating the relationship between these two metrics. Although the optimisation campaign successfully identified a highly selective reaction, the overall *ee* distribution was broad, with most reactions falling in the low *ee* range. In contrast, the conversion distribution was predominantly high across the campaign, and no discernible trend was found between conversion and *ee*, indicating that improvements in one metric did not necessarily correspond to changes in the other.

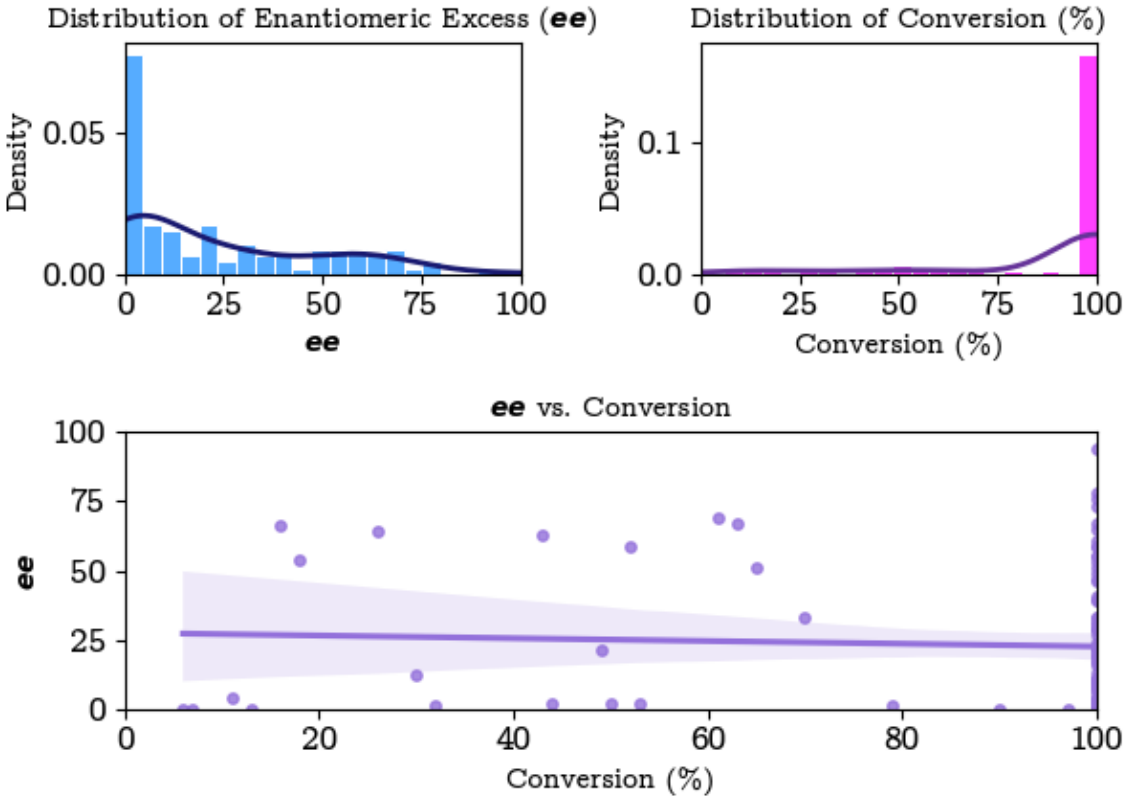


Figure 4.25: The distribution of *ee* and conversion across all 6 optimisation iterations. The bottom Scatter plot shows the relationship between *ee* and conversion (%) for the initial reactions. A linear regression is fitted to the data.

The skewness of the *ee* and conversion distributions was calculated for each optimisation iteration (**Figure 4.26**). Positive skewness values indicate distributions with a longer tail toward higher values, compared to a normal distribution, whereas negative values indicate a longer tail toward lower values. Across the campaign, *ee* distributions were consistently positively skewed, consistent with the majority of reactions giving low enantioselectivity and occasional high *ee* outliers. In contrast, conversion distributions were generally negatively skewed, consistent with most reactions achieving high conversion and a minority exhibiting low conversion.

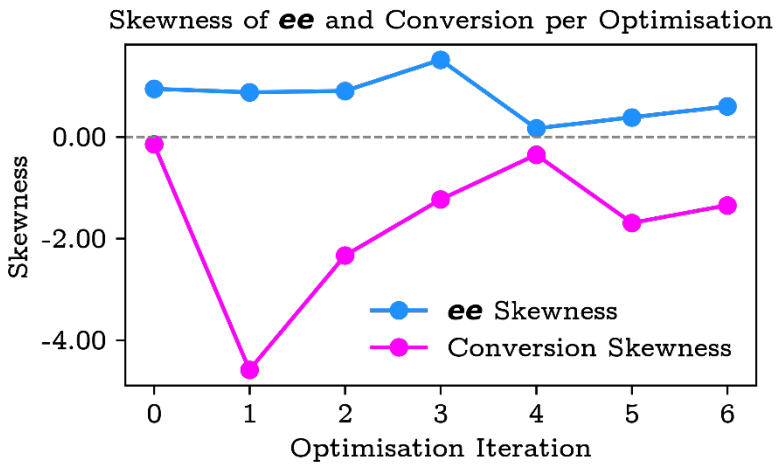


Figure 4.26: The skewness of *ee* and conversion across optimisation rounds.

Figure 4.27 displays the highest *ee* observed in each iteration, the best *ee* achieved so far, and the mean *ee* for each iteration. This figure highlights where improvements in the highest *ee* plateaued. The highest *ee* per iteration increased during the early stages, from the initial reactions (iteration 0) through optimisations one, two, and three, reflecting the BO algorithm's ability to identify increasingly selective conditions. However, a decrease was observed in Optimisation three before increasing through Optimisations four and five. Although the mean *ee* improved over most iterations, it remained far from optimal, as high selectivity was achieved in only a small number of experiments. This is typical of exploratory optimisation campaigns, where the search space contains large regions of suboptimal performance with only narrow zones of optimal conditions.

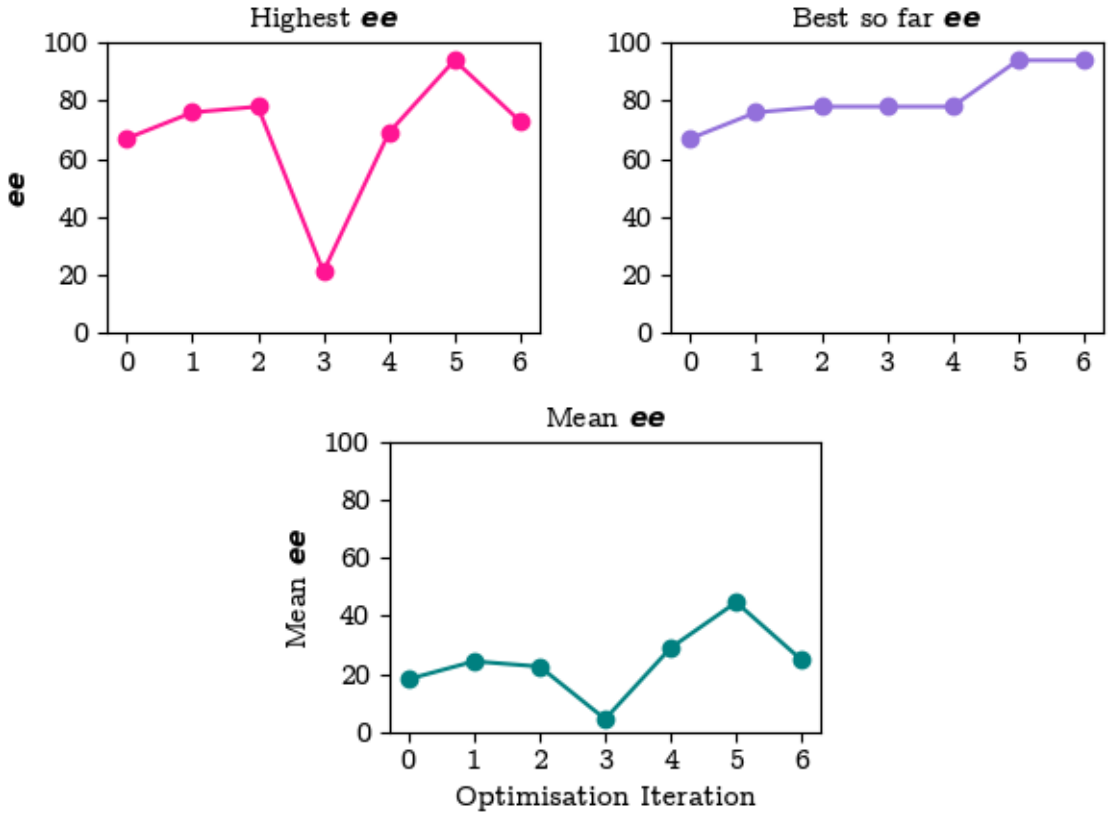
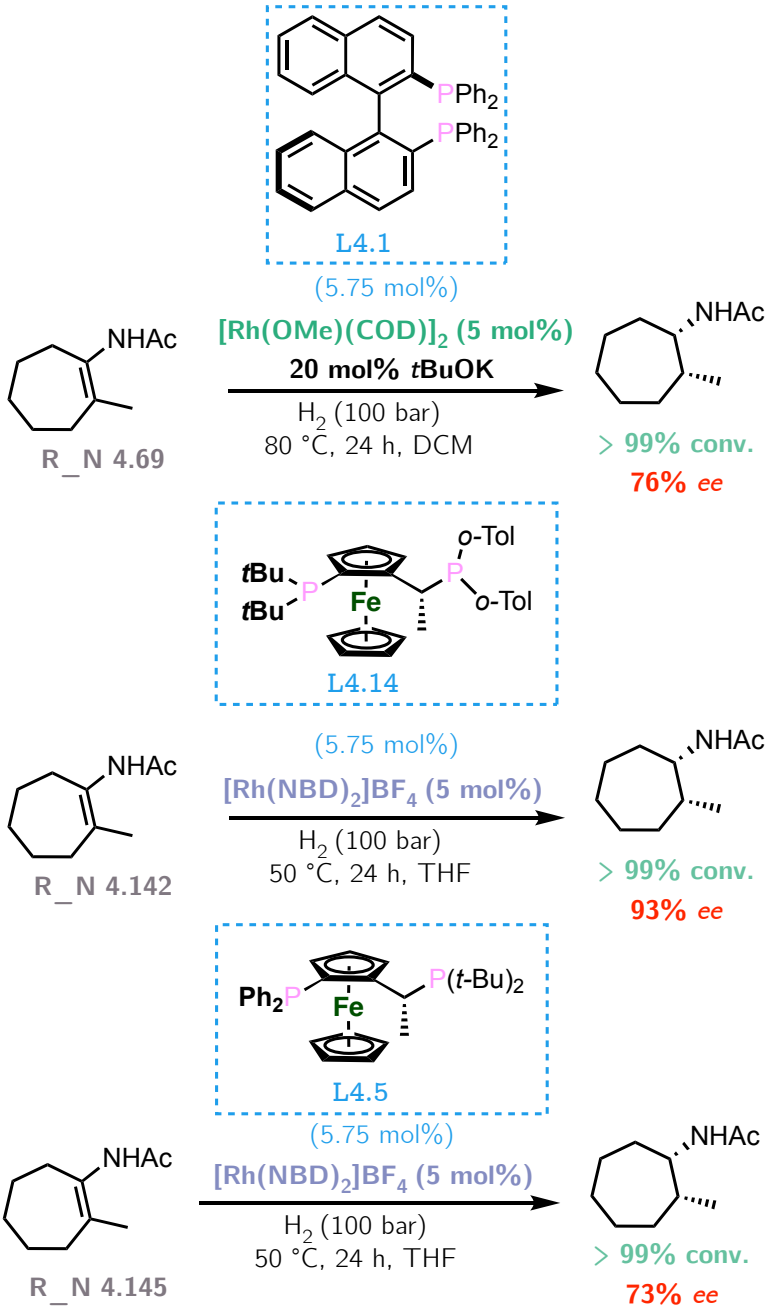


Figure 4.27: Performance metrics for the optimisation campaign. Comparison of three metrics for each optimisation iteration: (left, top) the highest *ee* achieved within each optimisation round, (right, top) the cumulative best *ee* observed up to that iteration; (centre, bottom) the mean *ee* across all reactions in that optimisation round.

The BO approach successfully identified a highly selective reaction (93% *ee*) using **L4.14** and optimised the reaction parameters for other chiral ligands, leading to improved selectivities for those systems (**Scheme 4.16**). As this project used commercially available chiral ligands, most had previously been applied in AH reactions of TSC=C. For example, **L4.5** was employed by Zhang in 2017 for the AH of β-aryl-substituted tetra-substituted enamides (**Scheme 4.4**),[203] and (*R*)-BINAP (**L4.1**) has been widely used throughout the AH literature.

In contrast, **L4.14** has been less extensively explored. Notably, it was used by a Merck team in 2017 for the synthesis of taranabant (**Scheme 4.3**),[297] and previously for the AH of α,β-unsaturated ketones with $Rh(COD)_2OTf$ and Zn(II) triflate at 70 bar H_2 and 70 °C.[306]



Scheme 4.16: The most enantioselective reaction found with from three different chiral ligands, reaction 4.69 located in Optimisation one, reaction 4.142 located in Optimisation five and reaction 4.145 located in Optimisation six.

4.4 CONCLUSIONS AND FUTURE WORK

This BO approach successfully identified a highly enantioselective reaction (93% *ee*) for the cyclic enamide substrate **S4.1**. In future work, the reaction conditions identified through this workflow could be applied to other cyclic enamides and assess performances across a diverse substrate scope.

The enantioselective reaction was identified within five rounds of optimisation, exploring only 8% of the defined reaction space. While this is a promising outcome, determining whether the approach is genuinely faster and more data-efficient than traditional screening would require direct comparison with results obtained by experienced experimental chemists. Likewise, performing several iterations of randomly selected reactions could help establish whether BO offers an advantage over conventional trial-and-error experimentation. For such comparisons to be fully accurate in identifying the most enantioselective reaction within the reaction space, all 1,782 possible reactions would need to be conducted, a labour-intensive task that, without HTE, lay beyond the scope of this project.

Future work could therefore focus on applying this approach to the optimisation of a different, but equally challenging, substrate to further assess its performance and generality.

4.5 EXPERIMENTAL

4.5.1 General Considerations

Unless otherwise stated all manipulations were carried out under an inert atmosphere (N_2 or Ar) using standard Schlenk techniques or employment of an MBraun glovebox. Glassware was dried in an oven at 180 °C overnight before use.

Unless otherwise stated, all reagents were purchased from Merck or Flurochem and used as purchased. Laboratory grade toluene and THF were dried over sodium/benzophenone and distilled prior to use. Ethanol (EtOH), DCM were heated at reflux over CaH and distilled onto activated 3Å molecular sieves. Room temperature (rt) refers to 298 K. Diisopropylamine was also dried over activated 3Å molecular sieves.

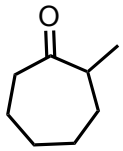
High pressure hydrogen reactions were conducted in a HEL cat24 high pressure reactor vessel.

^{1}H and $^{13}C\{^{1}H\}$ NMR spectra were recorded on a Brucker 400 MHz or 500 MHz spectrometers in $CDCl_3$ and ^{1}H and $^{13}C\{^{1}H\}$ NMR chemical shifts are reported relative to $CDCl_3$ at δ_H 7.26 ppm and δ_C 77.16 ppm. Coupling constants (J) are reported in Hertz (Hz). Multiplicities are indicated by: br s (broad singlet), s (singlet), d (doublet), t (triplet), q (quartet), p (pentet) and m (multiplet).

GC-MS analysis was obtained on a Shimadzu Qp2020 (EI) instrument equipped with a SH-Rxi-5MS column (25 m × 0.20 mm ID × 0.33 mm film).

4.5.2 Synthesis of *N*-(2-Methylcyclohept-1-en-1-yl)acetamide

2-methylcycloheptanone



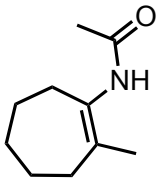
To a 250 mL flame dried Schlenk flask under an inert atmosphere n-butyllithium (*n*-BuLi, 1.6 M in hexane, 49 mL, 78.5 mmol) was added to a solution of diisopropylamine (11 mL, 78.5 mmol) in THF (100 mL) slowly at 0 °C. The solution was stirred for 30 min before being cooled to –78 °C. Freshly distilled cycloheptanone (9.2 mL, 78 mmol) was then added dropwise over 60 minutes and the resulting solution was stirred for a further 30 minutes. The solution was warmed to room temperature and stirred for 18 hours. The reaction was quenched with sat. aq. NH_4Cl and DCM, the organic layer was successively washed with aq. HCl (0.1 M) and sat. aq. $NaHCO_3$ and dried over $MgSO_4$.

The solvent was removed *in vacuo* and the product was purified by vacuum distillation at 1×10^{-1} mbar to give 2-methylcycloheptanone as a colourless oil (8.0 g, 81%). Analytical data is in accordance with the reported literature.[307]

^{1}H NMR (400 MHz, $CDCl_3$): δ 2.65 – 2.52 (m, 1H, CH), 2.52 – 2.42 (m, 2H, CH_2), 1.92 – 1.79 (m, 4H, CH_2), 1.72 – 1.52 (m, 4H, CH_2), 1.50 – 1.25 (m, 4H, CH_2), 1.04 (d, J = 6.9 Hz, 3H, CH_3).

^{13}C{^{1}H} NMR (101 MHz, $CDCl_3$) δ 216.9 (C=O), 46.7 (CH), 42.6 (CH_2), 33.4 (CH_2), 29.8 (CH_2), 28.8 (CH_2), 24.6 (CH_2), 17.6 (CH_3).

N-(2-methylcyclohept-1-en-1-yl)acetamide (S4.1)



To a 250 mL Schlenk flask under an inert atmosphere was charged 2-methylcycloheptanone (1.5 g, 11.9 mmol), hydroxylamine hydrochloride (1.15 g, 16.7 mmol), pyridine (2.4 mL, 29.8 mmol) and EtOH (100 mL). The resulting mixture was stirred at 60 ° C for 4 hours. The solvent was removed under vacuum and the mixture extracted with EtOAc. The organic layers were combined and volatiles removed at reduced pressure to give 2-methyl-cycloheptanone oxime.

To a 250 mL Schlenk flask under an inert atmosphere, iron powder (35.7 mmol, 2.0 g), acetic acid (71.4 mmol, 3.6 mL) and acetic anhydride (35.7 mmol, 2.8 mL) were added. The resulting mixture was stirred at 130 ° C for 6 hours to form iron(II) acetate in-situ.

2-methyl-cycloheptanone oxime was charged into the iron (II) acetate slowly, the resulting mixture was stirred at 60 ° C for 4 hours. The reaction mixture was then extracted with ethyl acetate and passed through celite to remove the metal precipitate. Volatiles were removed at reduced pressure and purified by silica gel chromatography (petroleum ether/ethyl acetate, 4:1), to give the desired product N-(2-methylcyclohept-1-en-1-yl)acetamide as a white solid (538 mg, 25% yield).
Analytical data is in accordance with reported literature.[4]

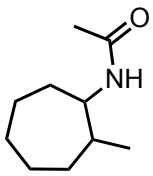
^{1}H NMR (500 MHz, $CDCl_3$): δ 2.40 – 2.28 (m, 2H), 2.19 – 2.08 (m, 2H), 2.01 (s, 3H, CH_3), 1.79 – 1.67 (m, 2H), 1.65 (s, 3H, CH_3), 1.60 – 1.53 (m, 2H), 1.52 – 1.45 (m, 2H).

^{13}C{^{1}H} NMR (126 MHz, $CDCl_3$): δ 168.3 (C=O), 132.5 (C=C), 131.9 (C=C), 34.4 (CH_2), 33.5 (CH_2), 32.1 (CH_2), 26.2 (CH_2), 25.9 (CH_2), 23.6 (CH_3), 20.6 (CH_3).

EI-MS: m/z 167.0 $[M]^+$

4.5.3 Synthesis of *N*-(-2-Methylcycloheptyl)acetamide

N-(-2-methylcycloheptyl)acetamide (P4.1)



The racemate N-(-2-methylcycloheptyl)acetamide was formed following procedure.

To 4 mL vial N-(2-methylcyclohept-1-en-1-yl)acetamide (0.1 mmol, 17 mg), Pd/C 10 wt% (0.01 mmol, 2 mg) and 500 uL ethyl acetate was charged. The vial was placed in high pressure autoclave (HEL cat24 high pressure reactor vessel). The autoclave was pressurised to 50 bar H_2 and the reaction was left stirring for 18 hours. After carefully depressurising the autoclave the reaction was passed through celite and the volatiles were subsequently removed at reduced pressure. Purification by column chromatography (petroleum ether/ethyl acetate, 4:1) obtained N-(-2-methylcycloheptyl)acetamide as a white solid (90% yield, 16 mg)
Analytical data is in accordance with reported literature for the (R,S/S,R) diastereomer.[4]

^{1}H NMR (500 MHz, $CDCl_3$): δ 5.43 (brs, 1H, NH), 4.21 – 4.03 (m, 1H, CH), 1.98 (s, 3H, CH_3), 1.98 – 1.93 (m, 1H), 1.72 – 1.46 (m, 8H, CH_2), 1.32 – 1.25 (m, 1H, CH), 0.89 (d, J = 7.0 Hz, 3H, C-CH_3).

^{13}C{^{1}H} NMR (126 MHz, $CDCl_3$): δ 169.3, 52.7, 36.5, 32.0, 28.1, 25.7, 23.9, 23.7, 17.7.

EI-MS: m/z 169.0 $[M]^+$

GC-MS: (19.51 major, 20.33 minor)

4.5.4 General Procedure for the Asymmetric Hydrogenation Reactions of *N*-(-2-Methylcycloheptyl)acetamide

In a nitrogen-filled glovebox a 4 mL vial equipped with a magnetic stirrer bar was charged the indicated pre-catalyst (5 mol%, 0.005 mmol), chiral ligand (5.75 mol%, 0.006 mmol) and solvent (0.5 mL). The resultant mixture was stirred for 30 min before N-(-2-methylcycloheptyl)acetamide (0.1 mmol, 17 mg) was added. The charged vial was placed into a stainless-steel autoclave (HEL cat24 high pressure reactor vessel). The autoclave was then purged with H_2 three times and charged to the indicated pressure and temperature. The mixture was stirred for 24 h before it was cooled to room temperature and depressurised. The resultant mixture was directly passed through a short plug of silica with EtOAc/petreoleum ether to give the desired product, evidenced by ^{1}H NMR. The conversion and *ee* of the product was determined by GC with a chiral stationary phase.

The analytical data of the enantiopure product was in accordance with that of the racemate in **4.5.3**.

5 Conclusion

The AH of tetra-substituted olefins remains one of the most challenging transformations in homogeneous catalysis due to the severe steric congestion of these substrates and the strong dependence of enantioselectivity on subtle substrate-catalyst interactions. Despite major advances in ligand and catalyst design, optimisation of these reactions typically relies on extensive experimental screening and expert chemical intuition. The overarching aim of this thesis was to investigate whether ML-based strategies could offer a complementary and more efficient framework for reaction outcome prediction and optimisation in this sensitive reaction class.

In this work, supervised ML models were developed to predict *ee* for AH reactions of tetra-substituted olefins using literature data. Both regression and classification approaches demonstrated that meaningful structure-reactivity relationships can be learned from existing experimental datasets, achieving prediction errors below 10% *ee* MAE and classification accuracies exceeding 80%.

Application of the trained models to novel experimental reactions highlighted both their promise and current limitations. While performance was better for reactions similar to those represented in the training data, predictive accuracy declined for low-*ee* reactions. This behaviour reflects a fundamental limitation of supervised learning when applied to experimental chemistry, where model generalisability is constrained by dataset size, diversity, and reporting bias. In the context of AH, the strong skew of literature data toward successful reactions remains a significant barrier to the production of efficient predictive models.

To address these challenges, BO was explored as an alternative strategy for experimental optimisation. By iteratively guiding experiments using probabilistic surrogate models, BO enabled efficient exploration of reaction space and the identification of highly enantioselective conditions using substantially fewer experiments than conventional screening approaches. This demonstrated the value of closed loop, data-driven experimentation for reaction classes where data is scarce and experimental costs are high.

Taken together, the findings of this thesis indicate that for this reaction class ML is most effective when applied as a decision support tool rather than a standalone predictive solution. When integrated with experimental workflows, data-driven approaches can reduce unproductive screening, prioritise promising regions of reaction space, and accelerate optimisation campaigns. While challenges remain, particularly in data availability and generalisability, this work demonstrates that ML-guided strategies offer a practical and impactful route toward more efficient development of AH reactions of tetra-substituted olefins.

6 References

1. M. Feher and J. M. Schmidt, *J. Chem. Inf. Model.*, 2003, **43**, 218-227.

2. B. M. Trost, J. D. Knopf and C. S. Brindle, *Chem. Rev.*, 2016, **116**, 15035-15088.

3. S. Kraft, K. Ryan and R. B. Kargbo, *J. Am. Chem. Soc.*, 2017, **139**, 11630-11641.

4. C. Li, F. Wan, Y. Chen, H. Peng, W. Tang, S. Yu, J. C. McWilliams, J. Mustakis, L. Samp and R. J. Maguire, *Angew. Chem. Int. Ed.*, 2019, **58**, 13573-13583.

5. L. A. Saudan, *Acc. Chem. Res.*, 2007, **40**, 1309-1319.

6. M. Takahashi, N. Suzuki and T. Ishikawa, *J. Org. Chem. Res.*, 2013, **78**, 3250-3261.

7. Y. Crameri, J. Foricher, U. Hengartner, C.-J. Jenny, F. Kienzle, H. Ramuz, M. Scalone, M. Schlageter, R. Schmid and S. Wang, *Chimia*, 1997, **51**, 303-303.

8. R. Bigler, K. A. Mack, J. Shen, P. Tosatti, C. Han, S. Bachmann, H. Zhang, M. Scalone, A. Pfaltz and S. E. Denmark, *Angew. Chem.*, 2020, **132**, 2866-2871.

9. K. Garber, *Nat. Biotechnol.*, 2013, **31**, 3-5.

10. M. Nilsson, M. Hämäläinen, M. Ivarsson, J. Gottfries, Y. Xue, S. Hansson, R. Isaksson and T. Fex, *J. Med. Chem.*, 2009, **52**, 2708-2715.

11. R. S. Greenberg, *J. Agric. Food. Chem.*, 1981, **29**, 856-860.

12. J. E. Frampton and M. P. Curran, *Drugs*, 2007, **67**, 1767-1792.

13. J. Tack and P. Janssen, *Expert Opin. Emerg. Drugs*, 2011, **16**, 283-292.

14. F. Lovering, J. Bikker and C. Humblet, *J. Med. Chem.*, 2009, **52**, 6752-6756.

15. T. Lei, F. Chen, H. Liu, H. Sun, Y. Kang, D. Li, Y. Li and T. Hou, *Mol. Pharmaceutics*, 2017, **14**, 2407-2421.

16. M. S. Taylor and E. N. Jacobsen, *Proc. Natl. Acad. Sci.*, 2004, **101**, 5368-5373.

17. S. J. Mohan, E. C. Mohan and M. R. Yamsani, *Int. J. Pharm. Sci. Nanotechnol.*, 2008, **1**, 309-316.

18. L. A. Nguyen, H. He and C. Pham-Huy, *Int. J. Biomed. Sci.*, 2006, **2**, 85-100.

19. N. Vargesson, *Birth Defects Res., Part C*, 2015, **105**, 140-156.

20. S.-A. Lim, *Ann. Acad. Med*, 2006, **35**, 274-278.

21. W. S. Knowles, M. J. Sabacky and B. D. Vineyard, *Ann. N. Y. Acad. Sci.*, 1973, **214**, 119-124.

22. W. Knowles, M. Sabacky and B. Vineyard, *Acc. Chem. Res.*, 1968, **40**, 1238-1239.

23. R. Noyori, T. Ohkuma, M. Kitamura, H. Takaya, N. Sayo, H. Kumobayashi and S. Akutagawa, *J. Am. Chem. Soc.*, 1987, **109**, 5856-5858.

24. W. S. Knowles, *Angew. Chem. Int. Ed.*, 2002, **41**, 1998-2007.

25. L. C. Xu, S. Q. Zhang, X. Li, M. J. Tang, P. P. Xie and X. Hong, *Angew. Chem.*, 2021, **133**, 22986-22993.

26. Z. Tu, T. Stuyver and C. W. Coley, *Chem. Sci.*, 2023, **14**, 226-244.

27. J. W. Sin, S. L. Chau, R. P. Burwood, K. Püntener, R. Bigler and P. Schwaller, *Nat. Commun.*, 2025, **16**, 6464.

28. H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2018, **4**, 1465-1476.

29. C. Benhaim, L. Bouchard, G. Pelletier, J. Sellstedt, L. Kristofova and S. Daigneault, *Org. Lett.*, 2010, **12**, 2008-2011.

30. J. García-Serna, R. Piñero-Hernanz and D. Durán-Martín, *Catal. Today*, 2022, **387**, 237 – 243.

31. X. Cui and K. Burgess, *Chem. Rev.*, 2005, **105**, 3272-3296.

32. M. Biosca, P. de la Cruz-Sánchez, J. Faiges, J. Margalef, E. Salomó, A. Riera, X. Verdaguer, J. Ferré, F. Maseras and M. Besora, *ACS Catal.*, 2023, **13**, 3020-3035.

33. M. V. Troutman, D. H. Appella and S. L. Buchwald, *J. Am. Chem. Soc.*, 1999, **121**, 4916-4917.

34. M. G. Schrems, E. Neumann and A. Pfaltz, *Angew. Chem. Int. Ed.*, 2007, **46**, 8274-8276.

35. J. Sprinz and G. Helmchen, *Tetrahedron Lett.*, 1993, **34**, 1769-1772.

36. C. A. Busacca, B. Qu, N. Grět, K. R. Fandrick, A. K. Saha, M. Marsini, D. Reeves, N. Haddad, M. Eriksson and J. P. Wu, *Adv. Synth. Catal.*, 2013, **355**, 1455-1463.

37. Z. Zhang, F. Yang, J.-J. Fu, Y.-H. Shen, W. He and W.-D. Zhang, *RSC Adv.*, 2016, **6**, 65885-65888.

38. Z. Zhang, J. Wang, J. Li, F. Yang, G. Liu, W. Tang, W. He, J.-J. Fu, Y.-H. Shen and A. Li, *J. Am. Chem. Soc.*, 2017, **139**, 5558-5567.

39. W. Tang, B. Qu, A. G. Capacci, S. Rodriguez, X. Wei, N. Haddad, B. Narayanan, S. Ma, N. Grinberg and N. K. Yee, *Org. Lett.*, 2010, **12**, 176-179.

40. M. Biosca, O. Pàmies and M. Diéguez, *Catal. Sci. Technol.*, 2020, **10**, 613-624.

41. M. J. Burk, M. F. Gross and J. P. Martinez, *J. Am. Chem. Soc.*, 1995, **117**, 9375-9376.

42. G. J. Roff, R. C. Lloyd and N. J. Turner, *J. Am. Chem. Soc.*, 2004, **126**, 4098-4099.

43. G. Hoge, H.-P. Wu, W. S. Kissel, D. A. Pflum, D. J. Greene and J. Bao, *J. Am. Chem. Soc.*, 2004, **126**, 5966-5967.

44. C. Molinaro, J. P. Scott, M. Shevlin, C. Wise, A. Menard, A. Gibb, E. M. Junker and D. Lieberman, *J. Am. Chem. Soc.*, 2015, **137**, 999-1006.

45. S. Ponra, W. Rabten, J. Yang, H. Wu, S. Kerdphon and P. G. Andersson, *J. Am. Chem. Soc.*, 2018, **140**, 13878-13883.

46. J. Halpern, *Science*, 1982, **217**, 401-407.

47. J. M. Brown, *Chem. Soc. Rev.*, 1993, **22**, 25-41.

48. C. R. Landis and T. W. Brauch, *Inorg. Chim. Acta*, 1998, **270**, 285-297.

49. C. R. Landis and J. Halpern, *J. Am. Chem. Soc.*, 1987, **109**, 1746-1754.

50. I. D. Gridnev, N. Higashi, K. Asakura and T. Imamoto, *J. Am. Chem. Soc.*, 2000, **122**, 7183-7194.

51. I. D. Gridnev, N. Higashi and T. Imamoto, *J. Am. Chem. Soc.*, 2000, **122**, 10486-10487.

52. K. Källström, C. Hedberg, P. Brandt, A. Bayer and P. G. Andersson, *J. Am. Chem. Soc.*, 2004, **126**, 14308-14309.

53. P. Brandt, C. Hedberg and P. G. Andersson, *Chem. Eur. J.*, 2003, **9**, 339-347.

54. X. Cui, Y. Fan, M. B. Hall and K. Burgess, *Chem. Eur. J.*, 2005, **11**, 6859-6868.

55. J. Mazuela, P.-O. Norrby, P. G. Andersson, O. Pamies and M. Dieguez, *J. Am. Chem. Soc.*, 2011, **133**, 13634-13645.

56. T. L. Church, T. Rasmussen and P. G. Andersson, *Organometallics*, 2010, **29**, 6769-6781.

57. S. Gruber and A. Pfaltz, *Angew. Chem.*, 2014, **126**, 1927-1931.

58. C. Borras, M. Biosca, O. Pamies and M. Dieguez, *Organometallics*, 2015, **34**, 5321-5334.

59. O. Pàmies, J. Zheng, J. Faiges and P. G. Andersson, *Adv. Synth. Catal.*, 2021, **68**, 135-203.

60. J. J. Verendel, O. Pamies, M. Dieguez and P. G. Andersson, *Chem. Rev.*, 2014, **114**, 2130-2169.

61. K. Källström, I. Munslow and P. G. Andersson, *Chem. Eur. J.*, 2006, **12**, 3194-3200.

62. S. Gruber, *Organometallics*, 2016, **35**, 699-705.

63. Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis and T. F. Jaramillo, *ACS Catal.*, 2017, **7**, 6600-6608.

64. K. Tran and Z. W. Ulissi, *Nat. Catal.*, 2018, **1**, 696-703.

65. B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360-365.

66. S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, *Nat. Commun.*, 2019, **10**, 539.

67. P. V. Balachandran, B. Kowalski, A. Sehirlioglu and T. Lookman, *Nat. Commun.*, 2018, **9**, 1668.

68. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186-190.

69. J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377-381.

70. G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.

71. Y. Zhuo, A. Mansouri Tehrani, A. O. Oliynyk, A. C. Duke and J. Brgoch, *Nat. Commun.*, 2018, **9**, 4377.

72. J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725-732.

73. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.

74. J. R. Kitchin, *Nat. Catal.*, 2018, **1**, 230-232.

75. M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255-260.

76. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547-555.

77. F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 872.

78. Z. Zhou, X. Li and R. N. Zare, *ACS Cent. Sci.*, 2017, **3**, 1337-1344.

79. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268-276.

80. S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2016, **55**, 5904-5937.

81. B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103-1113.

82. P. S. Gromski, A. B. Henson, J. M. Granda and L. Cronin, *Nature Rev. Chem.*, 2019, **3**, 119-128.

83. M. H. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604-610.

84. M. H. Segler and M. P. Waller, *Chem. Eur. J.*, 2017, **23**, 5966-5971.

85. L.-Y. Chen and Y.-P. Li, *Beilstein J. Org. Chem.*, 2024, **20**, 2476-2492.

86. L. Fiedler, K. Shah, M. Bussmann and A. Cangi, *Phys. Rev. Mater.*, 2022, **6**, 040301.

87. S. I. Simak, E. K. Delczeg-Czirjak and O. Eriksson, *Sci. Rep.*, 2025, **15**, 17212.

88. Y. Elbaz and M. Caspary Toroker, *Sci. Rep.*, 2024, **14**, 12150.

89. O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schutt, A. Tkatchenko and K.-R. Muller, *Chem. Rev.*, 2021, **121**, 10142-10186.

90. A. S. Anker, K. T. Butler, R. Selvan and K. M. Jensen, *Chem. Sci.*, 2023, **14**, 14003-14019.

91. P. Klukowski, R. Riek and P. Güntert, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2025, **148-149**, 101575.

92. J. Contreras and T. Bocklitz, *Pflügers Archiv-European Journal of Physiology*, 2025, **477**, 603-615.

93. F. Strieth-Kalthoff, F. Sandfort, M. H. Segler and F. Glorius, *Chem. Soc. Rev.*, 2020, **49**, 6154-6168.

94. S. Consuegra-Jiménez, C. Tovio-Gracia and R. Vivas-Reyes, *Results Chem.*, 2025, **17**, 102517.

95. A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields and A. G. Doyle, *Acc. Chem. Res.*, 2021, **54**, 1856-1865.

96. J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Muller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816-9872.

97. A. Tkatchenko, *Nat. Commun.*, 2020, **11**, 4125.

98. F. Strieth - Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, *Angew. Chem. Int. Ed.*, 2022, **61**, e202204647.

99. I. V. Tetko, O. Engkvist, U. Koch, J. L. Reymond and H. Chen, *Mol. Inf.*, 2016, **35**, 615-621.

100. P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73-76.

101. M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle and N. V. Chawla, *Chem. Sci.*, 2023, **14**, 4997-5005.

102. W. Chen, K. Yang, Z. Yu, Y. Shi and C. P. Chen, *Artif. Intell. Rev.*, 2024, **57**, 137.

103. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *Artif. Intell. Res.*, 2002, **16**, 321-357.

104. H. He, Y. Bai, E. A. Garcia and S. Li,ADASYN: Adaptive synthetic sampling approach for imbalanced learning, 2008.

105. A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.

106. S. Callaghan, *Patterns*, 2021, **2**, 100221.

107. J. R. Schmink, A. Bellomo and S. Berritt, *Aldrichimica Acta*, 2013, **46**, 71-80.

108. Y.-F. Shi, Z.-X. Yang, S. Ma, P.-L. Kang, C. Shang, P. Hu and Z.-P. Liu, *Engineering*, 2023, **27**, 70-83.

109. B. Huang, G. F. von Rudorff and O. A. von Lilienfeld, *Science*, 2023, **381**, 170-175.

110. T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve and D. J. Griffin, *J. Med. Chem.*, 2020, **63**, 8667-8682.

111. B. Karlik, M. F. Yilmaz, M. Ozdemir, C. T. Yavuz and Y. Danisman, *Plasmonics*, 2021, **16**, 147-155.

112. L. David, A. Thakkar, R. Mercado and O. Engkvist, *J. Cheminf.*, 2020, **12**, 56.

113. F. Weinhold, C. Landis and E. Glendening, *Int. Rev. Phys. Chem.*, 2016, **35**, 399-440.

114. S. Singh and R. B. Sunoj, *Acc. Chem. Res.*, 2023, **56**, 402-412.

115. G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler and M. Ceriotti, *Chem. Phys.*, 2018, **148**.

116. A. Capecchi, D. Probst and J.-L. Reymond, *J. Cheminf.*, 2020, **12**, 1-15.

117. D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742-754.

118. S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi, *J. Cheminf.*, 2015, **7**, 1-34.

119. A. T. Balaban, *J. Chem. Inf. Model.*, 1985, **25**, 334-343.

120. D. Weininger, *J. Chem. Inf. Model.*, 1988, **28**, 31-36.

121. D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Model.*, 1989, **29**, 97-101.

122. D. S. Wigh, J. M. Goodman and A. A. Lapkin, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2022, **12**, e1603.

123. N. M. O'Boyle, *J. Cheminf.*, 2012, **4**, 22.

124. RDKit (version 2025.03.6): Open-source cheminformatics, https://www.rdkit.org (accessed 3 October 2025).

125. N. Schneider, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2111-2120.

126. M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys and A. Vaitkus, *J. Cheminf.*, 2018, **10**, 23.

127. H. L. Morgan, *J. Chem. Doc*, 1965, **5**, 107-113.

128. S. A. Wildman and G. M. Crippen, *J. Chem. Inf. Model.*, 1999, **39**, 868-873.

129. J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219-3228.

130. A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313-2323.

131. L. B. De Amorim, G. D. Cavalcanti and R. M. Cruz, *Appl. Soft Comput.*, 2023, **133**, 109924.

132. S.-C. Li, H. Wu, A. Menon, K. A. Spiekermann, Y.-P. Li and W. H. Green, *J. Am. Chem. Soc.*, 2024, **146**, 23103-23120.

133. A. Pomberger, A. A. P. McCarthy, A. Khan, S. Sung, C. J. Taylor, M. J. Gaunt, L. Colwell, D. Walz and A. A. Lapkin, *React. Chem. Eng.*, 2022, **7**, 1368-1379.

134. D. M. Hawkins, *J. Chem. Inf. Model.*, 2004, **44**, 1-12.

135. R. Kumar and A. K. Singh, *Npj Comput. Mater.*, 2021, **7**, 197.

136. S. J. Guesné, T. Hanser, S. Werner, S. Boobier and S. Scott, *J. Cheminf.*, 2024, **16**, 43.

137. H. A. Duy and T. Srisongkram, *J. Cheminf.*, 2025, **17**, 39.

138. L. Breiman, *Mach. Learn.*, 2001, **45**, 5-32.

139. A. Bender, N. Schneider, M. Segler, W. Patrick Walters, O. Engkvist and T. Rodrigues, *Nature Rev. Chem.*, 2022, **6**, 428-442.

140. S. Käser, L. I. Vazquez-Salazar, M. Meuwly and K. Töpfer, *Digital Discovery*, 2023, **2**, 28-58.

141. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434-443.

142. P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572-1583.

143. P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Mach. Learn.: Sci. Technol.,* 2021, **2**, 015016.

144. J. Guo, B. Ranković and P. Schwaller, *chimia*, 2023, **77**, 31-31.

145. D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, Hoboken, NJ, 6th edn, 2021.

146. R. Tibshirani, *J. R. Stat. Soc. Series B Stat. Methodol.*, 1996, **58**, 267-288.

147. A. E. Hoerl and R. W. Kennard, *Technometrics*, 1970, **12**, 55-67.

148. S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *PNAS*, 2020, **117**, 1339-1345.

149. A. F. Zahrt and S. E. Denmark, *Tetrahedron*, 2019, **75**, 1841-1851.

150. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *J. Chem. Inf. Model.*, 2003, **43**, 1947-1958.

151. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *J. Chem. Inf. Model.*, 2003, **43**, 1947-1958.

152. L. Breiman, *Mach. Learn.*, 1996, **24**, 123-140.

153. P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3-42.

154. L. Breiman, *Neural Comput.*, 1999, **11**, 1493-1517.

155. J. H. Friedman, *Ann. Stat.*, 2001, 1189-1232.

156. A. Natekin and A. Knoll, *Front. Neurorobot.*, 2013, **7**, 21.

157. R. Akbani, S. Kwek and N. Japkowicz, Applying support vector machines to imbalanced datasets, in *Proceedings of the 15th European Conference on Machine Learning (ECML 2004)*, Springer, Berlin, 2004, pp. 39–50.

158. T. Hofmann, B. Schölkopf and A. J. Smola, *Ann. Stat.*, 2008, **36**, 1171-1220.

159. M. Das, A. Ghosh and R. B. Sunoj, *J. Comput. Chem.*, 2024, **45**, 1160-1176.

160. B. Li, S. Su, C. Zhu, J. Lin, X. Hu, L. Su, Z. Yu, K. Liao and H. Chen, *J. Cheminf.*, 2023, **15**, 72.

161. S.-W. Li, L.-C. Xu, C. Zhang, S.-Q. Zhang and X. Hong, *Nat. Commun.*, 2023, **14**, 3569.

162. E. Heid and W. H. Green, *J. Chem. Inf. Model.*, 2022, **62**, 2101-2110.

163. A. F. Zahrt, S. V. Athavale and S. E. Denmark, *Chem. Rev.*, 2019, **120**, 1620-1689.

164. A. Tomberg, M. J. Johansson and P.-O. Norrby, *J. Org. Chem. Res.*, 2018, **84**, 4695-4703.

165. J. Aires-de-Sousa and J. Gasteiger, *J. Comb. Chem.*, 2005, **7**, 298-301.

166. P. J. Donoghue, P. Helquist, P.-O. Norrby and O. Wiest, *J. Am. Chem. Soc.*, 2009, **131**, 410-411.

167. Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2021, **12**, 2198-2208.

168. L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. Oliveira, S.-W. Li, L. Ackermann and X. Hong, *Nat. Synth.*, 2023, **2**, 321-330.

169. S. Shilpa, G. Kashyap and R. B. Sunoj, *J. Phys. Chem. A.*, 2023, **127**, 8253-8271.

170. A. V. Kalikadien, C. Valsecchi, R. van Putten, T. Maes, M. Muuronen, N. Dyubankova, L. Lefort and E. A. Pidko, *Chem. Sci.*, 2024, **15**, 13618-13630.

171. A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist and R. H. Munday, *Nat. Catal.*, 2019, **2**, 41-45.

172. Y. Guan and S. E. Wheeler, *Angew. Chem. Int. Ed.*, 2017, **56**, 9101-9105.

173. W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, *ACS Cent. Sci.*, 2021, **7**, 1622-1637.

174. G. dos Passos Gomes, R. Pollice and A. Aspuru-Guzik, *Trends Chem.*, 2021, **3**, 96-110.

175. F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379-1390.

176. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang and J. Cai, *Pattern recognit.*, 2018, **77**, 354-377.

177. R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087-2096.

178. P. C. Hawkins, *J. Chem. Inf. Model.*, 2017, **57**, 1747-1756.

179. W. Kohn and L. J. Sham, *Phys. Rev. Lett.*, 1965, **140**, A1133.

180. S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787-1799.

181. Y. Cong, Y. Zhai, X. Chen and H. Li, *Int. J. Chem. Sci.*, 2022, **23**, 13371.

182. S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989-2009.

183. C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652-1671.

184. Z. Li and H. A. Scheraga, *PNAS*, 1987, **84**, 6611-6615.

185. M. Saunders, *J. Am. Chem. Soc.*, 1987, **109**, 3150-3152.

186. A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024-10035.

187. M. Bursch, A. Hansen, P. Pracht, J. T. Kohn and S. Grimme, *Phys. Chem. Chem. Phys.*, 2021, **23**, 287-299.

188. P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169-7192.

189. S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 2847-2862.

190. P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde and S. Grimme, *J. Comput. Aided Mol. Des.*, 2018, **32**, 1139-1149.

191. S. Spicher and S. Grimme, *Angew. Chem. Int. Ed.*, 2020, **59**, 15665-15673.

192. J. G. Brandenburg, C. Bannwarth, A. Hansen and S. Grimme, *Chem. Phys.*, 2018, **148**, 064104.

193. P. Pracht and S. Grimme, *Chem. Sci.*, 2021, **12**, 6551-6568.

194. S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, *J. Phys. Chem. A.*, 2021, **125**, 4039-4054.

195. S. Axelrod and R. Gomez-Bombarelli, *Sci. Data*, 2022, **9**, 1-14.

196. S. Dohm, M. Bursch, A. Hansen and S. Grimme, *J. Chem. Theory Comput.*, 2020, **16**, 2002-2012.

197. T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario and M. S. Sigman, *J. Am. Chem. Soc.*, 2022, **144**, 1205-1217.

198. A. Schumacher, M. G. Schrems and A. Pfaltz, *Chem. Eur. J.*, 2011, **17**, 13502-13509.

199. X. Du, Y. Xiao, Y. Yang, Y. N. Duan, F. Li, Q. Hu, L. W. Chung, G. Q. Chen and X. Zhang, *Angew. Chem. Int. Ed.*, 2021, **60**, 11384-11390.

200. S. Kerdphon, S. Ponra, J. Yang, H. Wu, L. Eriksson and P. G. Andersson, *ACS Catal.*, 2019, **9**, 6169-6176.

201. S.-F. Zhu and Q.-L. Zhou, *Acc. Chem. Res.*, 2017, **50**, 988-1001.

202. M. Biosca, E. Salomó, P. de la Cruz-Sánchez, A. Riera, X. Verdaguer, O. Pàmies and M. Diéguez, *Org. Lett.*, 2019, **21**, 807-811.

203. X. Li, C. You, H. Yang, J. Che, P. Chen, Y. Yang, H. Lv and X. Zhang, *Adv. Synth. Catal.*, 2017, **359**, 597-602.

204. M. Ashfaq, R. Tabassum, M. Ahmad, N. Hassan, H. Oku and G. Rivera, *Med. Chem*, 2015, **5**, 295-309.

205. Z. Han, Y.-Q. Guan, G. Liu, R. Wang, X. Yin, Q. Zhao, H. Cong, X.-Q. Dong and X. Zhang, *Org. Lett.*, 2018, **20**, 6349-6353.

206. T. Kajiwara, T. Konishi and M. Yamano, *Catal. Sci. Technol.*, 2012, **2**, 2146-2152.

207. J. D. Armstrong III, K. K. Eng, J. L. Keller, R. M. Purick, F. W. Hartner Jr, W.-B. Choi, D. Askin and R. Volante, *Tetrahedron Lett.*, 1994, **35**, 3239-3242.

208. A. Ohashi, S. i. Kikuchi, M. Yasutake and T. Imamoto, *Eur. J. Org. Chem.*, 2002, **2002**, 2535-2546.

209. J. Meng, M. Gao, H. Lv and X. Zhang, *Org. Lett.*, 2015, **17**, 1842-1845.

210. S. Song, S.-F. Zhu, Y. Li and Q.-L. Zhou, *Org. Lett.*, 2013, **15**, 3722-3725.

211. L. Shi, Z. S. Ye, L. L. Cao, R. N. Guo, Y. Hu and Y. G. Zhou, *Angew. Chem. Int. Ed.*, 2012, **51**, 8286-8289.

212. Y.-Q. Guan, Z. Han, X. Li, C. You, X. Tan, H. Lv and X. Zhang, *Chem. Sci.*, 2019, **10**, 252-256.

213. M. Biosca, M. Magre, O. Pamies and M. Dieguez, *ACS Catal.*, 2018, **8**, 10316-10320.

214. C. Molinaro, E. M. Phillips, B. Xiang, E. Milczek, M. Shevlin, J. Balsells, S. Ceglia, J. Chen, L. Chen and Q. Chen, *J. Org. Chem. Res.*, 2019, **84**, 8006-8018.

215. M. Yamada, T. Kato, T. Ikemoto and M. Yamano, *Org. Process Res. Dev.*, 2021, **25**, 1074-1079.

216. Y.-Q. Guan, M. Gao, X. Deng, H. Lv and X. Zhang, *Chem. Commun.*, 2017, **53**, 8136-8139.

217. Y.-T. Liu, J.-Q. Chen, L.-P. Li, X.-Y. Shao, J.-H. Xie and Q.-L. Zhou, *Org. Lett.*, 2017, **19**, 3231-3234.

218. M. Ma, G. Hou, J. Wang and X. Zhang, *Tetrahedron: Asymmetry*, 2011, **22**, 506-511.

219. M. Christensen, A. Nolting, M. Shevlin, M. Weisel, P. E. Maligres, J. Lee, R. K. Orr, C. W. Plummer, M. T. Tudge and L.-C. Campeau, *J. Org. Chem. Res.*, 2016, **81**, 824-830.

220. J. R. Calvin, M. O. Frederick, D. L. Laird, J. R. Remacle and S. A. May, *Org. Lett.*, 2012, **14**, 1038-1041.

221. S. W. Krska, J. V. Mitten, P. G. Dormer, D. Mowrey, F. Machrouhi, Y. Sun and T. D. Nelson, *Tetrahedron*, 2009, **65**, 8987-8994.

222. A. Stumpf, M. Reynolds, D. Sutherlin, S. Babu, E. Bappert, F. Spindler, M. Welch and J. Gaudino, *Adv. Synth. Catal.*, 2011, **353**, 3367-3372.

223. V. M. Shoba and J. M. Takacs, *J. Am. Chem. Soc.*, 2017, **139**, 5740-5743.

224. Q.-A. Chen, Z.-S. Ye, Y. Duan and Y.-G. Zhou, *Chem. Soc. Rev.*, 2013, **42**, 497-511.

225. L. Panella, A. M. Aleixandre, G. J. Kruidhof, J. Robertus, B. L. Feringa, J. G. de Vries and A. J. Minnaard, *J. Org. Chem. Res.*, 2006, **71**, 2026-2036.

226. Z. Zhang, N. A. Butt and W. Zhang, *Chem. Rev.*, 2016, **116**, 14769-14827.

227. X. Du, Y. Xiao, J.-M. Huang, Y. Zhang, Y.-N. Duan, H. Wang, C. Shi, G.-Q. Chen and X. Zhang, *Nat. Commun.*, 2020, **11**, 1-10.

228. H. Bernsmann, M. van den Berg, R. Hoen, A. J. Minnaard, G. Mehler, M. T. Reetz, J. G. De Vries and B. L. Feringa, *J. Org. Chem. Res.*, 2005, **70**, 943-951.

229. Y. Sawatsugawa, K. Tamura, N. Sano and T. Imamoto, *Org. Lett.*, 2019, **21**, 8874-8878.

230. M. Engman, J. S. Diesen, A. Paptchikhine and P. G. Andersson, *J. Am. Chem. Soc.*, 2007, **129**, 4536-4537.

231. T. L. Church and P. G. Andersson, *Coord. Chem. Rev.*, 2008, **252**, 513-531.

232. J. Wen, X. Fan, R. Tan, H.-C. Chien, Q. Zhou, L. W. Chung and X. Zhang, *Org. Lett.*, 2018, **20**, 2143-2147.

233. T. Touge and T. Arai, *J. Am. Chem. Soc.*, 2016, **138**, 11299-11305.

234. L. Eberhardt, D. Armspach, J. Harrowfield and D. Matt, *Chem. Soc. Rev.*, 2008, **37**, 839-864.

235. P. Kaukoranta, M. Engman, C. Hedberg, J. Bergquist and P. G. Andersson, *Adv. Synth. Catal.*, 2008, **350**, 1168-1176.

236. B. Zupančič, B. Mohar and M. Stephan, *Org. Lett.*, 2010, **12**, 3022-3025.

237. R. Hoen, M. van den Berg, H. Bernsmann, A. J. Minnaard, J. G. de Vries and B. L. Feringa, *Org. Lett.*, 2004, **6**, 1433-1436.

238. J. Meeuwissen, R. J. Detz, A. J. Sandee, B. de Bruin and J. N. Reek, *Dalton Trans.*, 2010, **39**, 1929-1931.

239. G. H. Bernardinelli, E. P. Kündig, P. Meier, A. Pfaltz, K. Radkowski, N. Zimmermann and M. Neuburger - Zehnder, *Helv. Chim. Acta*, 2001, **84**, 3233-3246.

240. D. G. Blackmond, A. Lightfoot, A. Pfaltz, T. Rosner, P. Schnider and N. Zimmermann, *Chirality*, 2000, **12**, 442-449.

241. B. Mohar and M. Stephan, *Adv. Synth. Catal.*, 2013, **355**, 594-600.

242. D. J. Wallace, K. R. Campos, C. S. Shultz, A. Klapars, D. Zewge, B. R. Crump, B. D. Phenix, J. C. McWilliams, S. Krska, Y. Sun, C.-y. Chen and F. Spindler, *Org. Process Res. Dev.*, 2009, **13**, 84-90.

243. G. S. Forman, T. Ohkuma, W. P. Hems and R. Noyori, *Tetrahedron Lett.*, 2000, **41**, 9471-9475.

244. P. Etayo and A. Vidal-Ferran, *Chem. Soc. Rev.*, 2013, **42**, 728-754.

245. P. Hannen, H. C. Militzer, E. M. Vogl and F. A. Rampf, *Chem. Commun.*, 2003, **9**, 2210-2211.

246. X.-B. Wang, D.-W. Wang, S.-M. Lu, C.-B. Yu and Y.-G. Zhou, *Tetrahedron: Asymmetry*, 2009, **20**, 1040-1045.

247. P. E. Maligres, S. W. Krska and G. R. Humphrey, *Org. Lett.*, 2004, **6**, 3147-3150.

248. G. Xu and S. R. Gilbertson, *Tetrahedron Lett.*, 2003, **44**, 953-955.

249. K. Bert, T. Noël, W. Kimpe, J. L. Goeman and J. Van der Eycken, *Org. Biomol. Chem.*, 2012, **10**, 8539-8550.

250. G. Luchini, J. Alegre-Requena, I. Funes-Ardoiz and R. S. Paton, *F1000research*, 2020, **9**, 291.

251. N. M. O'Boyle, C. Morley and G. R. Hutchison, *Chem. Cent. J.*, 2008, **2**, 1-7.

252. N. M. O'Boyle, A. L. Tenderholt and K. M. Langner, *J. Comput. Chem.*, 2008, **29**, 839-845.

253. L. Falivene, Z. Cao, A. Petta, L. Serra, A. Poater, R. Oliva, V. Scarano and L. Cavallo*, Nat. Chem.*, 2019, **11**, 872 – 879.

254. L. H. Hall and L. B. Kier, *Rev. Comput. Chem.*, 1991, 367-422.

255. S. C. Basak, G. D. Grunwald and S. Majumdar, *Mol2Net, Proceedings*, 2015, **1**, 1–10.

256. D. Bonchev and N. Trinajstić, *Chem. Phys.*, 1977, **67**, 4517-4533.

257. A. T. Balaban, *Chem. Phys. Lett.*, 1982, **89**, 399-404.

258. S. H. Bertz, *J. Am. Chem. Soc.*, 1981, **103**, 3599-3601.

259. Z. Wang, C. Li, B. Zheng, Q. Tan, Q. Wu, L. Liu, J. E. Nycz, T. Huang and T. Chen, *Org. Biomol. Chem.*, 2025, **23**, 7132-7137.

260. A. Singh, K. Fatima, A. Srivastava, S. Khwaja, D. Priya, A. Singh, G. Mahajan, S. Alam, A. K. Saxena and D. Mondhe, *Chem. Biol. Drug Des.*, 2016, **88**, 625-634.

261. R. Bansal, R. Singh, T. S. Dutta, Z. A. Dar and A. Bajpai, *Drug Discov. Today*, 2024, **29**, 104063.

262. M. H. Keylor, B. S. Matsuura and C. R. Stephenson, *Chem. Rev.*, 2015, **115**, 8976-9027.

263. A. S. Jadhav, Y. A. Pankhade, R. Hazra and R. V. Anand, *J. Org. Chem. Res.*, 2018, **83**, 10107-10119.

264. V. B. Kharitonov, S. A. Runikhina, Y. V. Nelyubina, D. V. Muratov, D. Chusov and D. A. Loginov, *Chem. Eur. J.*, 2021, **27**, 10903-10912.

265. C. M. Farr, A. M. Kazerouni, B. Park, C. D. Poff, J. Won, K. R. Sharp, M.-H. Baik and S. B. Blakey, *J. Am. Chem. Soc.*, 2020, **142**, 13996-14004.

266. K. Zhao and R. R. Knowles, *J. Am. Chem. Soc.*, 2022, **144**, 137-144.

267. S. Kainz, A. Brinkmann, W. Leitner and A. Pfaltz, *J. Am. Chem. Soc.*, 1999, **121**, 6421-6429.

268. F. Hase, L. M. Roch, C. Kreisbeck and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 1134-1145.

269. R.-R. Griffiths and J. M. Hernández-Lobato, *Chem. Sci.*, 2020, **11**, 577-586.

270. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li and R. Clowes, *Nature*, 2020, **583**, 237-241.

271. A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne and A. A. Lapkin, *Chem. Eng. J.*, 2018, **352**, 277-282.

272. D. M. Negoescu, P. I. Frazier and W. B. Powell, *INFORMS Journal on Computing*, 2011, **23**, 346-363.

273. B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89-96.

274. F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, M. Christensen, E. Liles, J. E. Hein and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 035021-035021.

275. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, *Proc. IEEE*, 2016, **104**, 148-175.

276. P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584-4601.

277. K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163-1175.

278. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006.

279. R. J. Hickman, J. Ruža, H. Tribukait, L. M. Roch and A. García-Durán, *React. Chem. Eng.*, 2023, **8**, 2284-2296.

280. K. Pearson, *Lond. Edinb. Dublin philos. mag.*, 1901, **2**, 559-572.

281. A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi and P. Nantermet, *Science*, 2015, **347**, 49-53.

282. M. Kondo, H. D. P. Wathsala, M. Sako, Y. Hanatani, K. Ishikawa, S. Hara, T. Takaai, T. Washio, S. Takizawa and H. Sasai, *Chem. Commun.*, 2020, **56**, 1259-1262.

283. F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch and A. Aspuru-Guzik, *Appl. Phys. Rev.*, 2021, **8**.

284. D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, 429-434.

285. R. Lee, *Chem. Ing. Tech.*, 2019, **91**, 191-200.

286. O. Mitsunobu and M. Yamada, *Bull. Chem. Soc. Jpn.*, 1967, **40**, 2380-2382.

287. B. Ranković, R.-R. Griffiths, H. B. Moss and P. Schwaller, *Digital Discovery*, 2024, **3**, 654-666.

288. K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, Chichester, UK, 2001.

289. D. N. Jumbam, R. A. Skilton, A. J. Parrott, R. A. Bourne and M. Poliakoff, *J. Flow Chem.*, 2012, **2**, 24-27.

290. J. S. Moore and K. F. Jensen, *Org. Process Res. Dev.*, 2012, **16**, 1409-1415.

291. B. E. Walker, J. H. Bannock, A. M. Nightingale and J. C. deMello, *React. Chem. Eng.*, 2017, **2**, 785-798.

292. J. A. G. Torres, S. H. Lau, P. Anchuri, J. M. Stevens, J. E. Tabora, J. Li, A. Borovika, R. P. Adams and A. G. Doyle, *J. Am. Chem. Soc.*, 2022, **144**, 19999-20007.

293. E. Bradford, A. M. Schweidtmann and A. Lapkin, *Journal of Global Optimization*, 2018, **71**, 407-438.

294. Y. Amar, Artur M. Schweidtmann, P. Deutsch, L. Cao and A. Lapkin, *Chem. Sci.*, 2019, **10**, 6697-6706.

295. F. Häse, L. M. Roch and A. Aspuru-Guzik, *Chem. Sci.*, 2018, **9**, 7642-7655.

296. C.-B. Yu, K. Gao, Q.-A. Chen, M.-W. Chen and Y.-G. Zhou, *Tetrahedron Lett.*, 2012, **53**, 2560-2563.

297. D. J. Wallace, K. R. Campos, C. S. Shultz, A. Klapars, D. Zewge, B. R. Crump, B. D. Phenix, J. C. McWilliams, S. Krska and Y. Sun, *Org. Process Res. Dev.*, 2009, **13**, 84-90.

298. F. W. Patureau, S. de Boer, M. Kuil, J. Meeuwissen, P.-A. R. Breuil, M. A. Siegler, A. L. Spek, A. J. Sandee, B. de Bruin and J. N. H. Reek, *J. Am. Chem. Soc.*, 2009, **131**, 6683-6685.

299. Z. Zhang, G. Zhu, Q. Jiang, D. Xiao and X. Zhang, *J. Org. Chem. Res.*, 1999, **64**, 1774-1775.

300. P. Dupau, C. Bruneau and P. H. Dixneuf, *Adv. Synth. Catal.*, 2001, **343**, 331-334.

301. W. Tang, S. Wu and X. Zhang, *J. Am. Chem. Soc.*, 2003, **125**, 9570-9571.

302. T. M. Fong, X.-M. Guan, D. J. Marsh, C.-P. Shen, D. S. Stribling, K. M. Rosko, J. Lao, H. Yu, Y. Feng and J. C. Xiao, *The Journal of pharmacology and experimental therapeutics*, 2007, **321**, 1013-1022.

303. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *JMLR*, 2011, **12**, 2825-2830.

304. S. Diwale, M. K. Eisner, C. Carpenter, W. Sun, G. C. Rutledge and R. D. Braatz, *Mol. Syst. Des. Eng.*, 2022, **7**, 622-636.

305. Y. Tang, W. Guo and J. Gao, Efficient Model Selection for Support Vector Machine with Gaussian Kernel Function, *Proc. IEEE Symp. Comput. Intell. Data Min. (CIDM 2009)*, IEEE, Nashville, TN, USA, 2009, pp. 40–45.

306. J. R. Calvin, M. O. Frederick, D. L. T. Laird, J. R. Remacle and S. A. May, *Org. Lett.*, 2012, **14**, 1038-1041.

307. A. Fürstner, M. Albert, J. Mlynarski, M. Matheu and E. DeClercq, *J. Am. Chem. Soc.*, 2003, **125**, 13132-13142.

7 Appendix

7.1 CHAPTER 2 APPENDIX

The dataset supporting this thesis, *Research data supporting Using Machine Learning to Optimise Asymmetric Hydrogenation Reactions of Tetra-substituted Olefins*, has been deposited in the University of Cambridge data repository (Apollo).

Table 7.1: The train and test MAE (% *ee*) for all the regression models with different input features.

Model	Train MAE (% *ee*)	Test MAE (% *ee*)
LLR(2D)	8.0 ± 0.2	15.4 ± 1.3
RR(2D)	5.6 ± 0.3	18.8 ± 0.9
RFR(2D)	16.0 ± 0.3	16.7 ± 1.2
ETR (2D)	8.2 ± 0.2	11.9 ± 1.3
GBR(2D)	9.5 ± 0.9	13.5 ± 0.9
LLR (3D)	16.7 ± 0.3	18.2 ± 0.9
RR(3D)	16.3 ± 0.3	18.3 ± 0.7
RFR(3D)	17.4 ± 0.3	18.3 ± 1.1
ETR (3D)	8.7 ± 0.2	13.7 ± 0.9
GBR(3D)	7.8 ± 0.3	14.4 ± 0.6
SVR(3D)	5.1 ± 0.3	12.7 ± 1.0
LLR(2D + 3D)	7.0 ± 0.2	15.4 ± 1.3
RR(2D + 3D)	10.2 ± 0.2	15.0 ± 1.2
RFR(2D + 3D)	16.7 ± 0.3	17.5 ± 1.1
ETR(2D + 3D)	6.4 ± 0.2	9.8 ± 1.2
GBR(2D + 3D)	8.7 ± 0.7	12.9 ± 0.9

Table 7.2: Optimum set of hyperparameters found for regression models with 2D input features.

Model	Hyperparameters	Values
LLR	α	0.001
RR	α	10
RFR	Max_features	0.8
	Max_samples	0.1
	Max_depth:	50
	n_estimators	800
GBR	Learning_rate	0.1
	Max_depth	8
	Max_features	0.5
	Min_samples_leaf	4
	Min_samples_split	10
	n_estimators	800
ETR	Max_features	0.5
	Min_samples_leaf	2
	Max_depth	None
	n_estimators	800

Table 7.3: Optimum set of hyperparameters found for regression models with 3D input features.

Model	Hyperparameters	Values
LLR	α	0.001
RR	α	10
RFR	Max_features	0.9
	Max_samples	0.1
	Max_depth:	2
	n_estimators	800
GBR	Learning_rate	0.05
	Max_depth	4
	Max_features	0.5
	n_estimators	300
ETR	Max_features	0.4
	Min_samples_leaf	4
	Max_depth	None
	n_estimators	800
SVR	kernel	rbf
	Gamma	scale
	C	4
	Epsilon	1e-06

Table 7.4: Optimum set of hyperparameters found for regression models with 2D + 3D input features.

Model	Hyperparameters	Values
LLR	α	0.001
RR	α	10
RFR	Max_features	0.8
	Max_samples	0.1
	Max_depth:	None
	n_estimators	800
GBR	Learning_rate	0.1
	Max_depth	8
	Max_features	0.5
	Min_samples_leaf	4
	Min_samples_split	10
	n_estimators	800
ETR	Max_features	0.4
	Min_samples_leaf	4
	Max_depth	None
	n_estimators	800

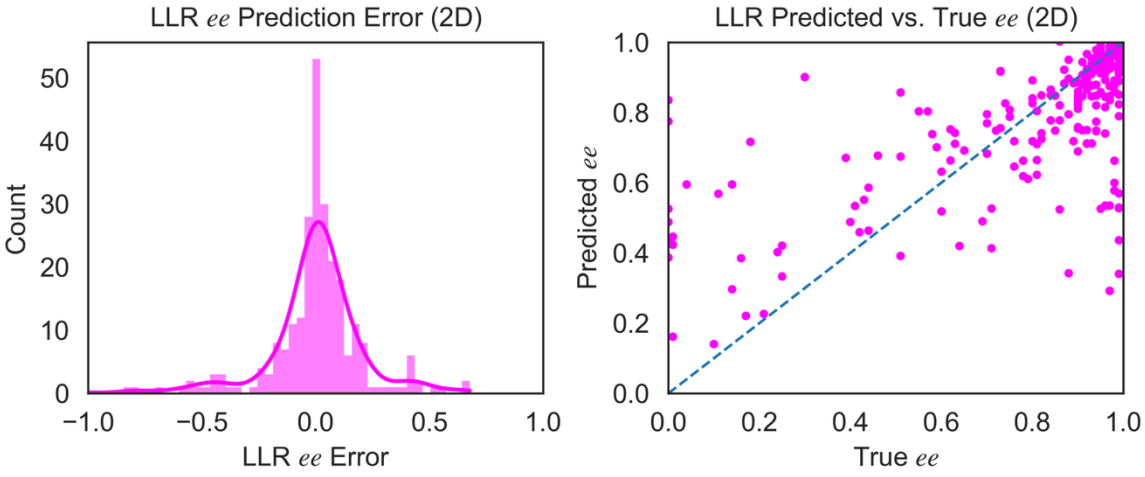


Figure 7.1: Performance of the LLR model for *ee* prediction across 2D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

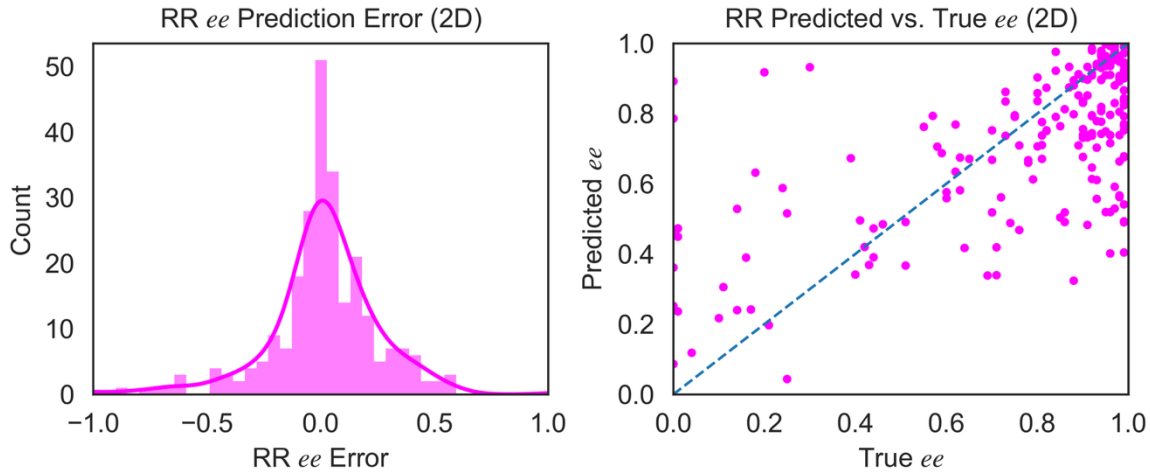


Figure 7.2: Performance of the RR model for *ee* prediction across 2D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

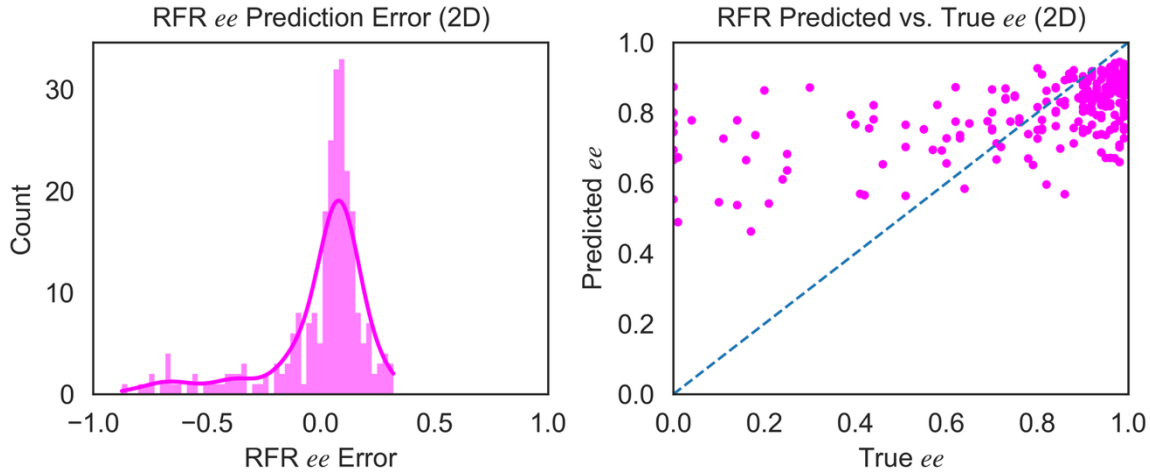


Figure 7.3: Performance of the RFR model for *ee* prediction across 2D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

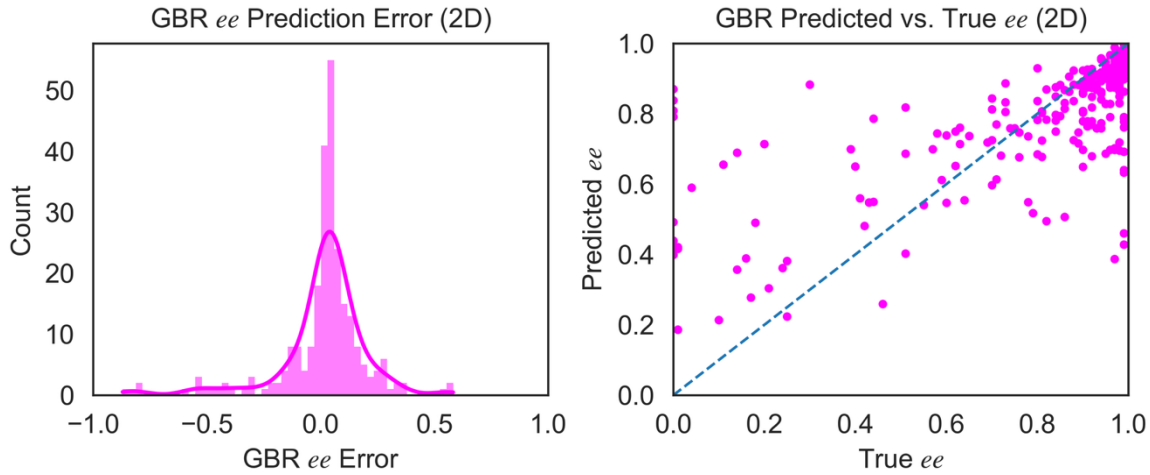


Figure 7.4: Performance of the GBR model for *ee* prediction across 2D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

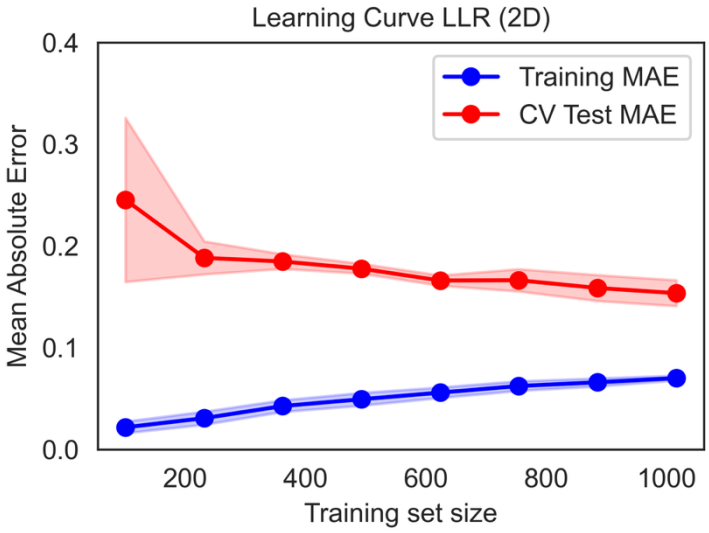


Figure 7.5: Learning curves for the LLR(2D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

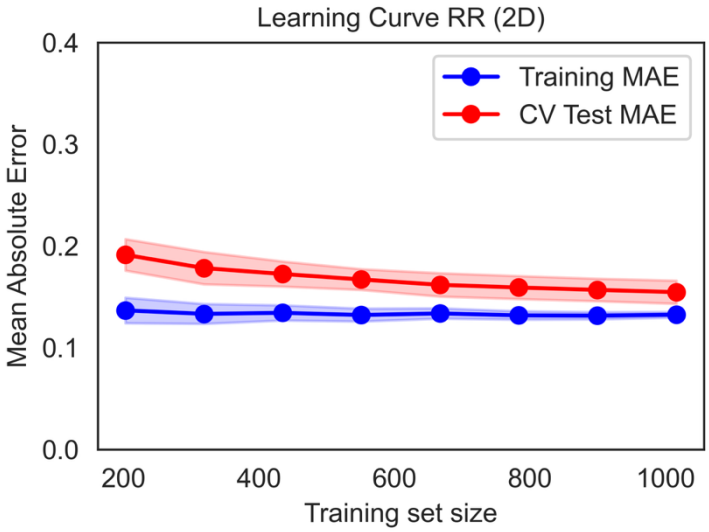


Figure 7.6: Learning curves for the RR(2D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

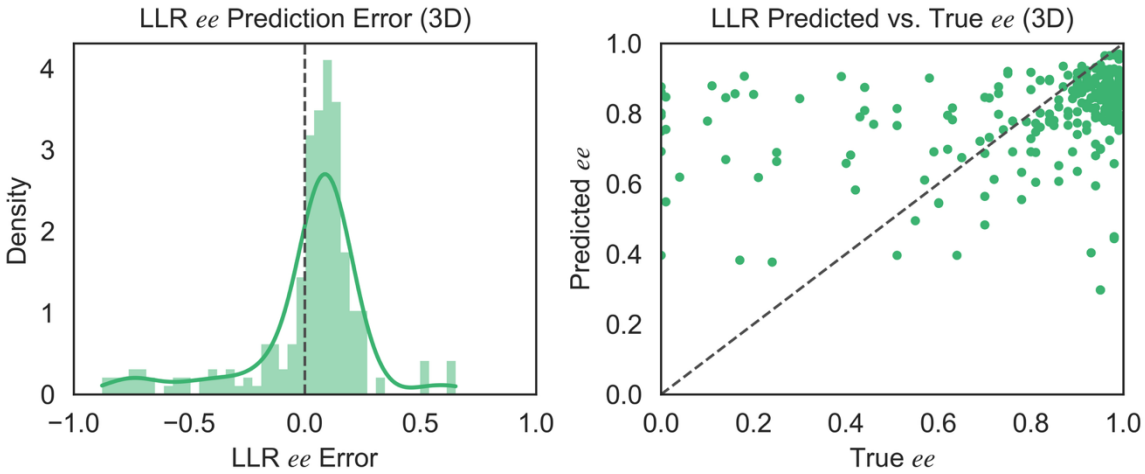


Figure 7.7: Performance of the LLR model for *ee* prediction across 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

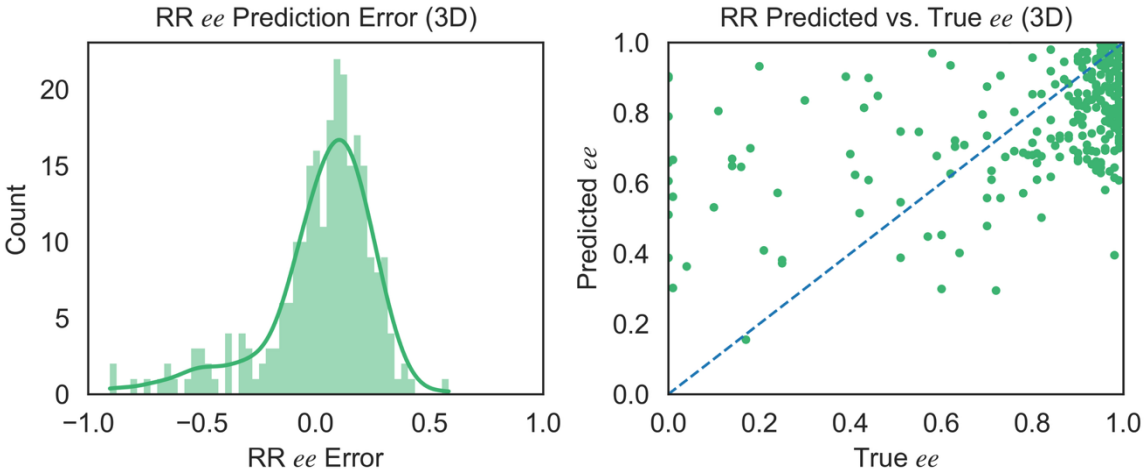


Figure 7.8: Performance of the RR model for *ee* prediction across 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

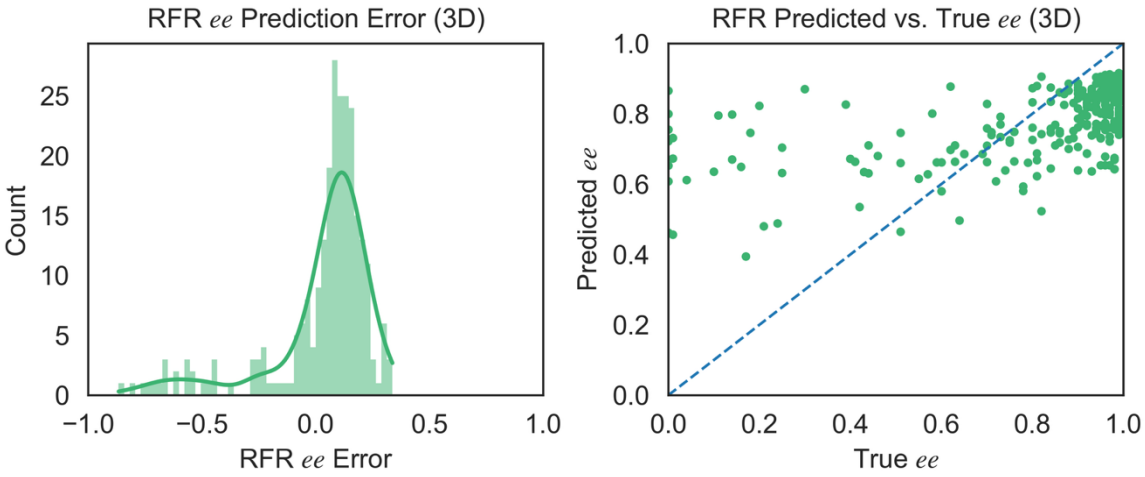


Figure 7.9: Performance of the RFR model for *ee* prediction across 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

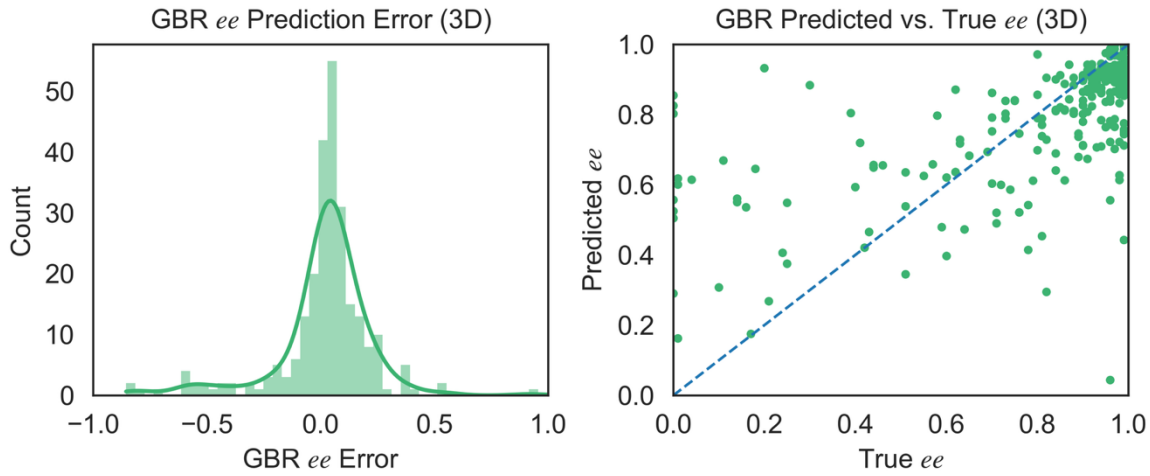


Figure 7.10: Performance of the GBR model for *ee* prediction across 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

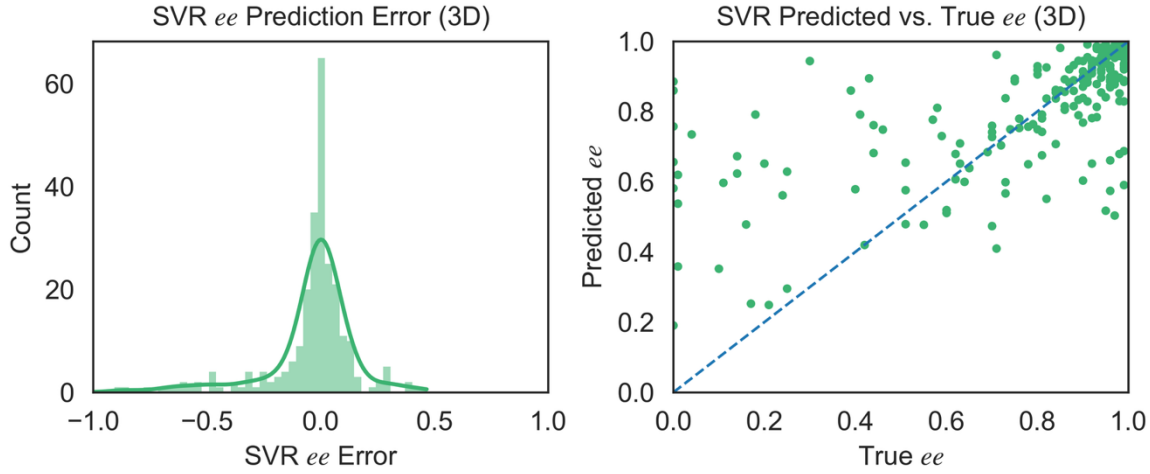


Figure 7.11: Performance of the SVR model for *ee* prediction across 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

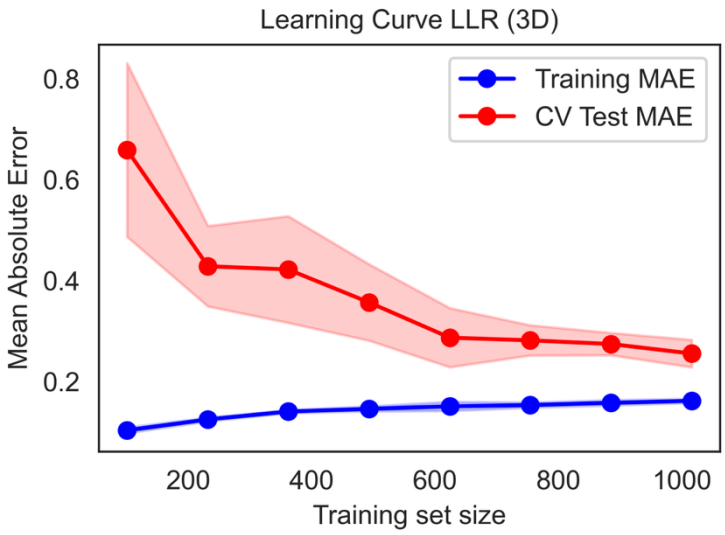


Figure 7.12: Learning curves for the LLR(3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

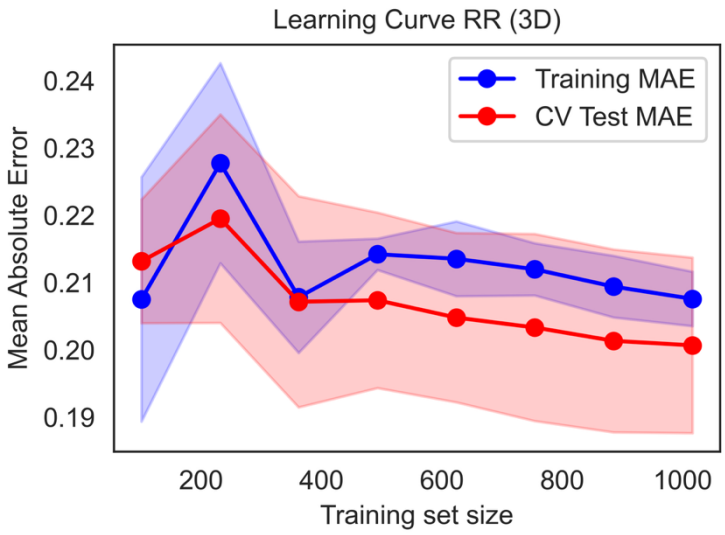


Figure 7.13: Learning curves for the RR(3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

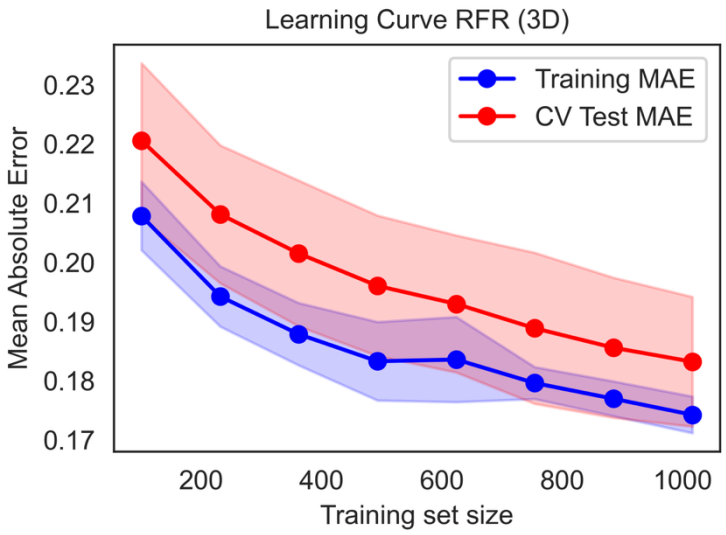


Figure 7.14: Learning curves for the RFR(3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

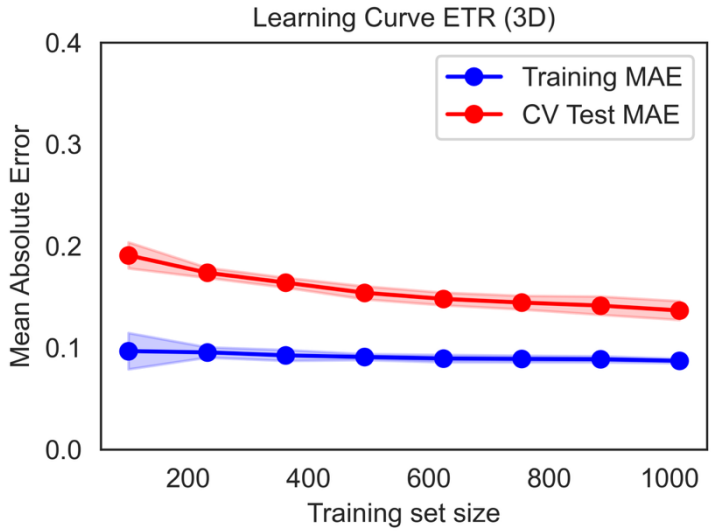


Figure 7.15: Learning curves for the ETR(3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

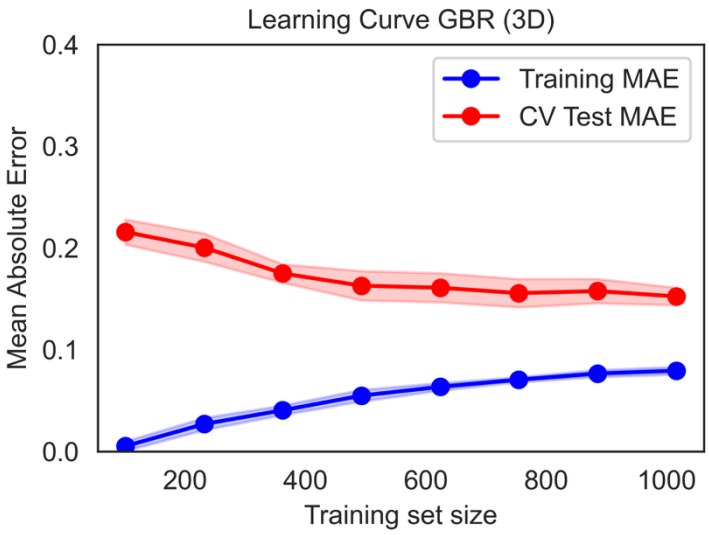


Figure 7.16: Learning curves for the GBR(3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

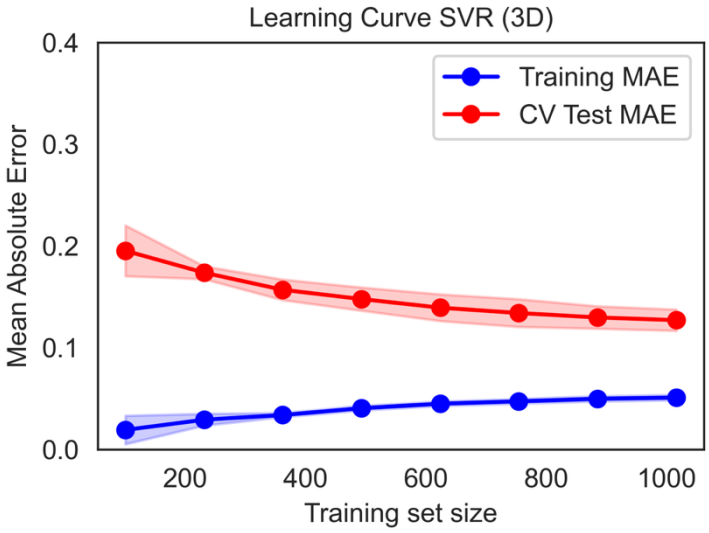


Figure 7.17: Learning curves for the SVR(3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

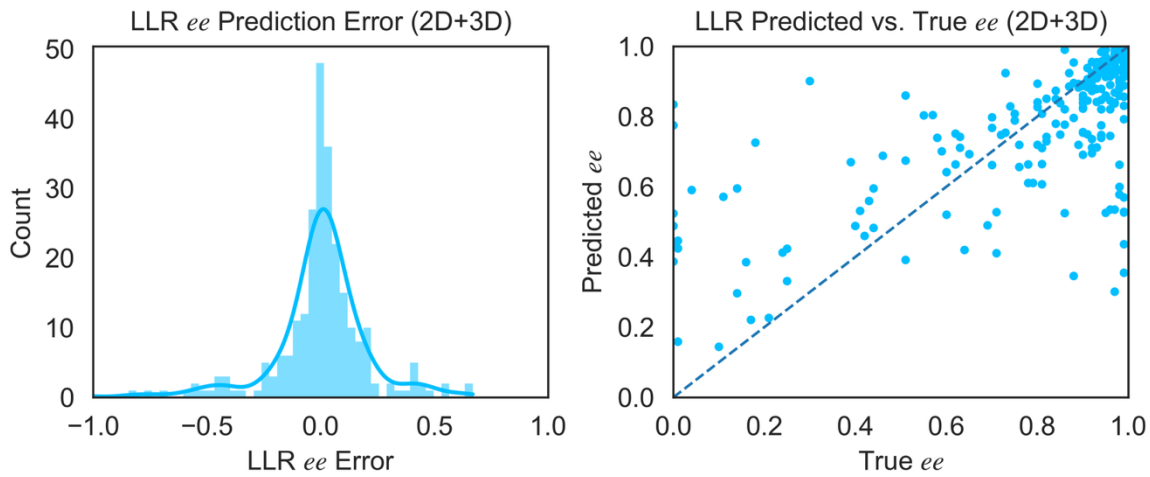


Figure 7.18: Performance of the LLR model for *ee* prediction across 2D + 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

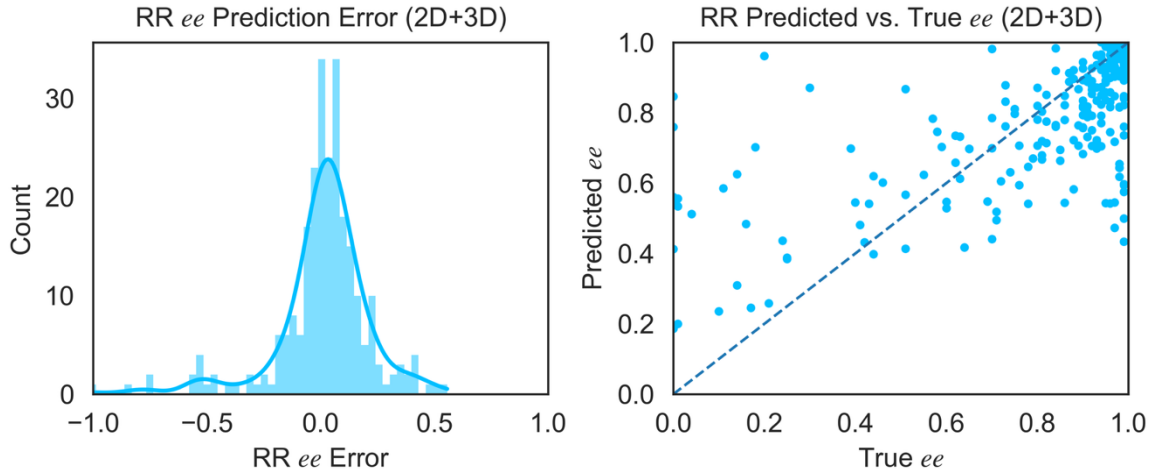


Figure 7.19: Performance of the RR model for *ee* prediction across 2D + 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

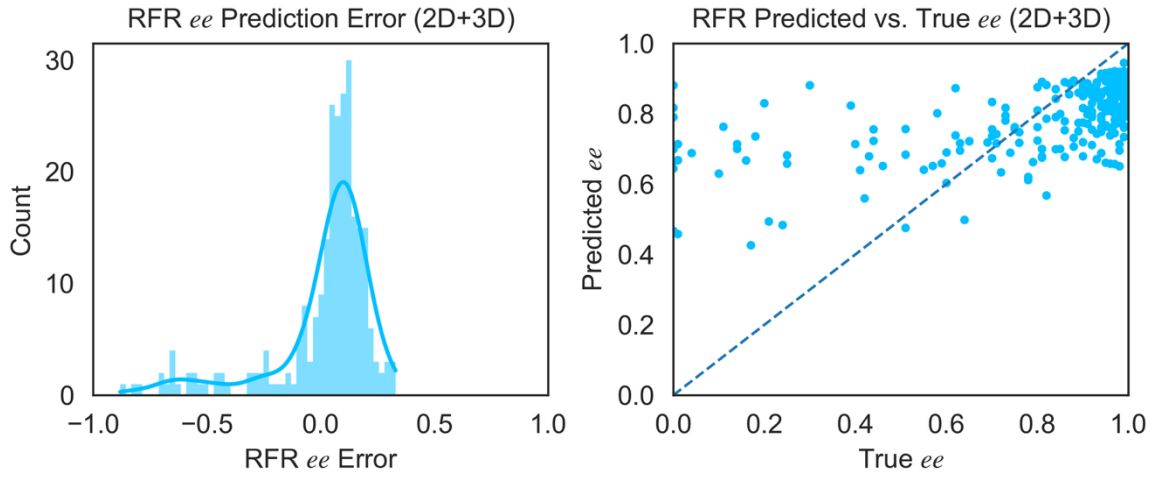


Figure 7.20: Performance of the RFR model for *ee* prediction across 2D + 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

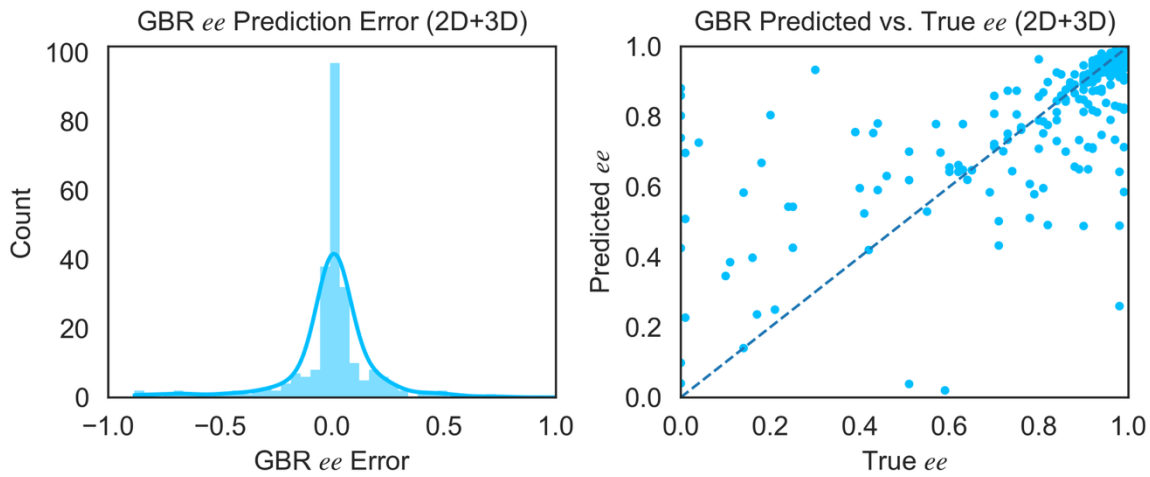


Figure 7.21: Performance of the GBR model for *ee* prediction across 2D + 3D input features. (Left) Distribution of prediction errors on the test set. (Right) scatter plot comparing predicted versus true *ee* values for the test set, with the dashed line indicating perfect prediction (y=x).

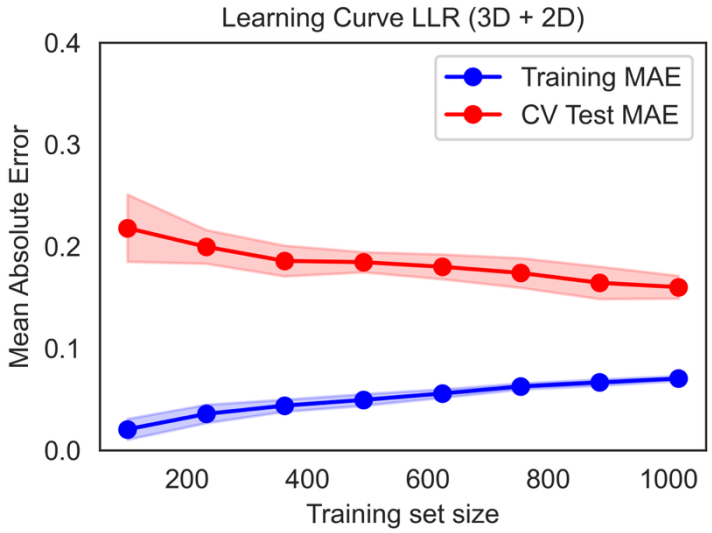


Figure 7.22: Learning curves for the LLR(2D + 3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

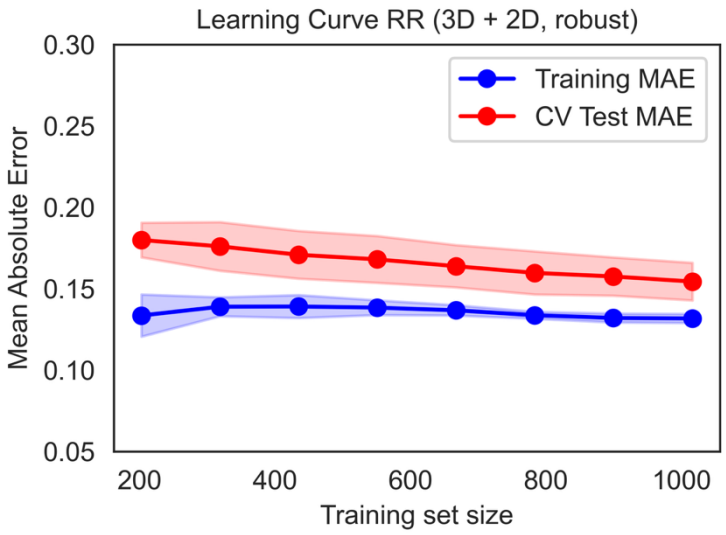
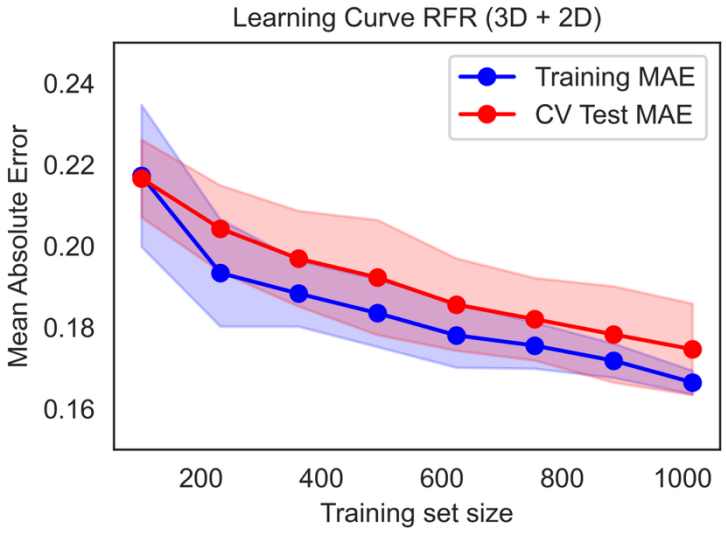


Figure 7.23: Learning curves for the RR(2D + 3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.



7.24: Learning curves for the RFR(2D + 3D) model. Blue markers show the CV training MAE and the red markers show the CV test MAE. The shaded bands denote the SD.

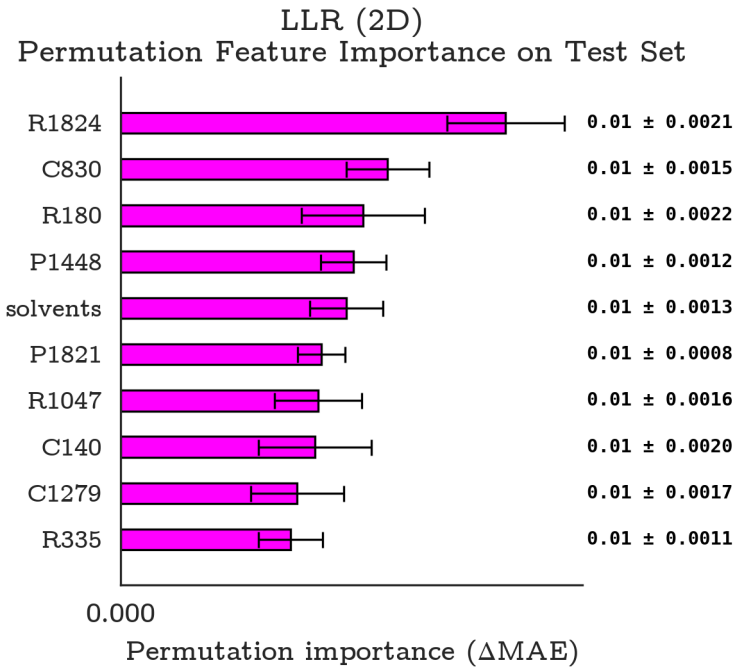


Figure 7.25: Permutation feature importance for the LLR(2D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

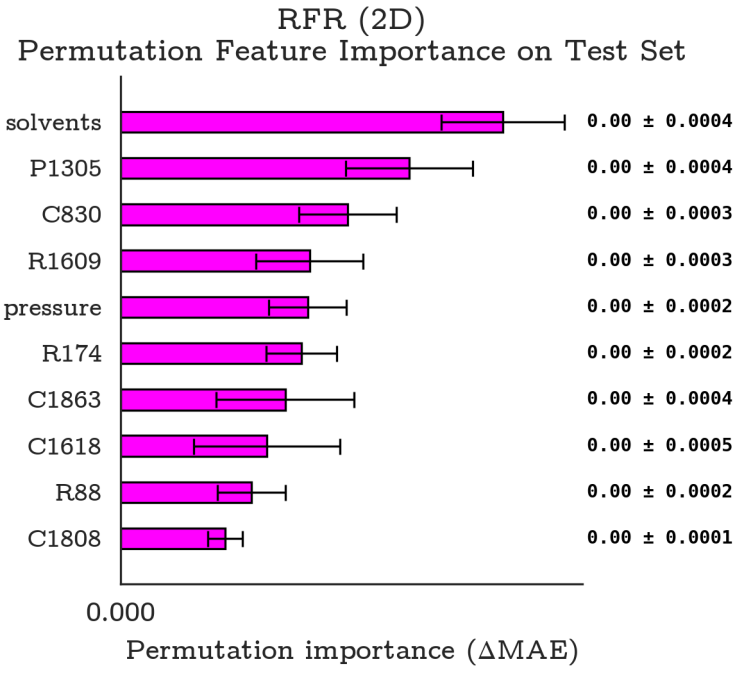


Figure 7.26: Permutation feature importance for the RFR(2D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

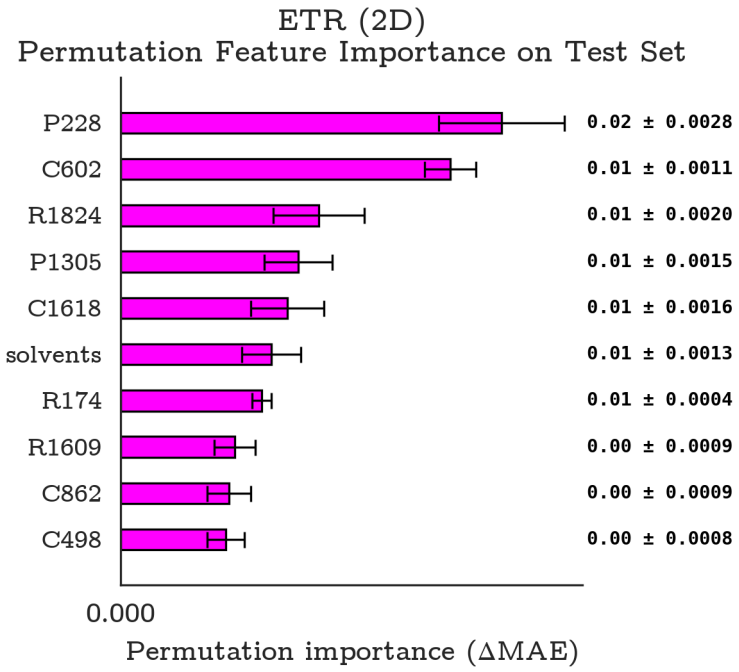


Figure 7.27: Permutation feature importance for the ETR(2D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

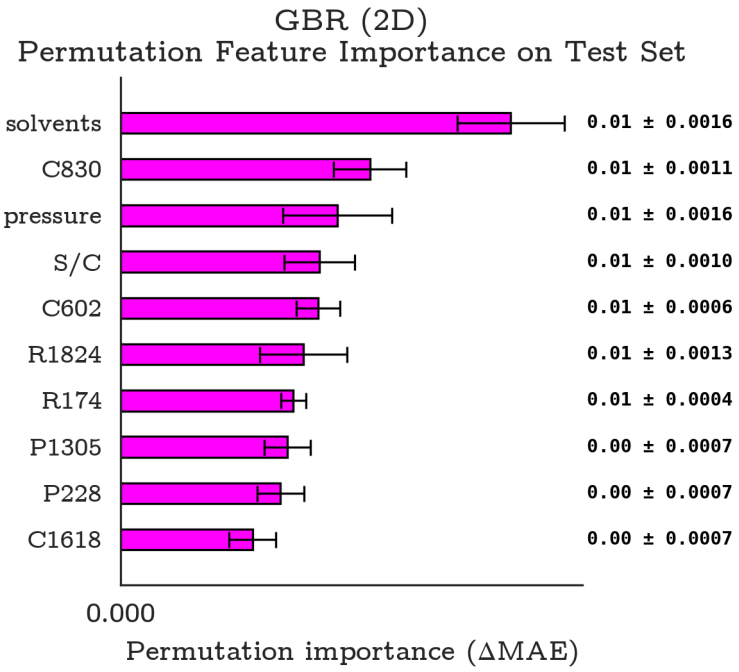


Figure 7.28: Permutation feature importance for the GBR(2D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

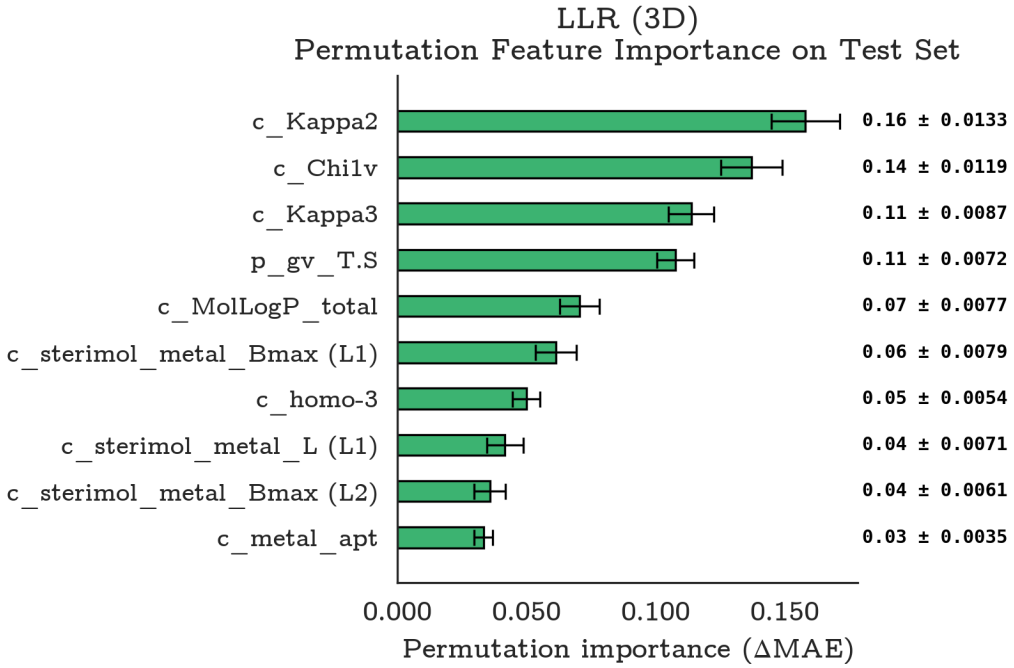


Figure 7.29: Permutation feature importance for the LLR(3D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

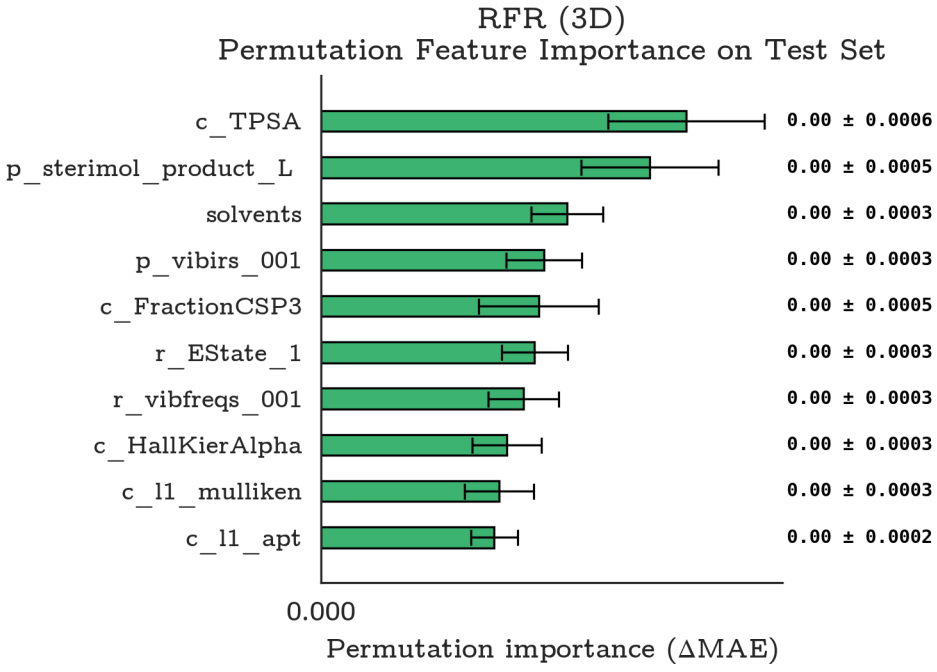


Figure 7.30: Permutation feature importance for the RFR(3D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

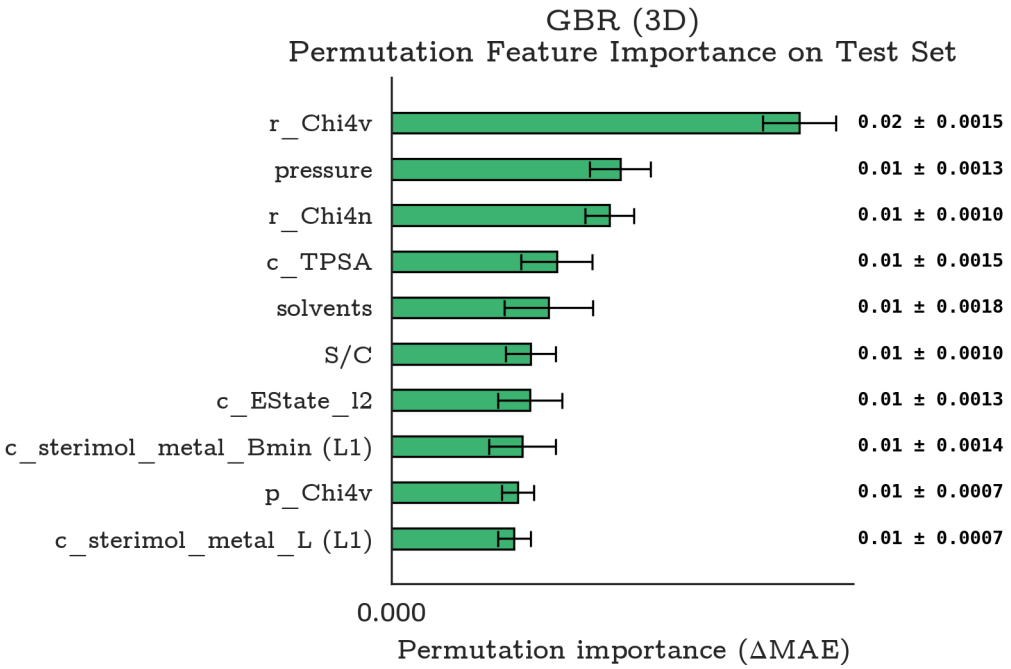


Figure 7.31: Permutation feature importance for the GBR(3D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

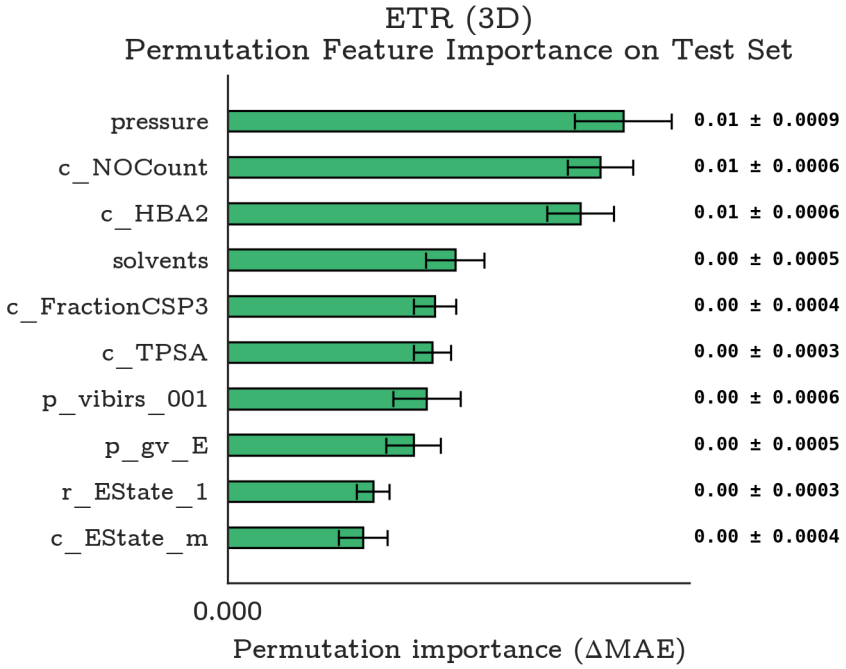


Figure 7.32: Permutation feature importance for the ETR(3D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

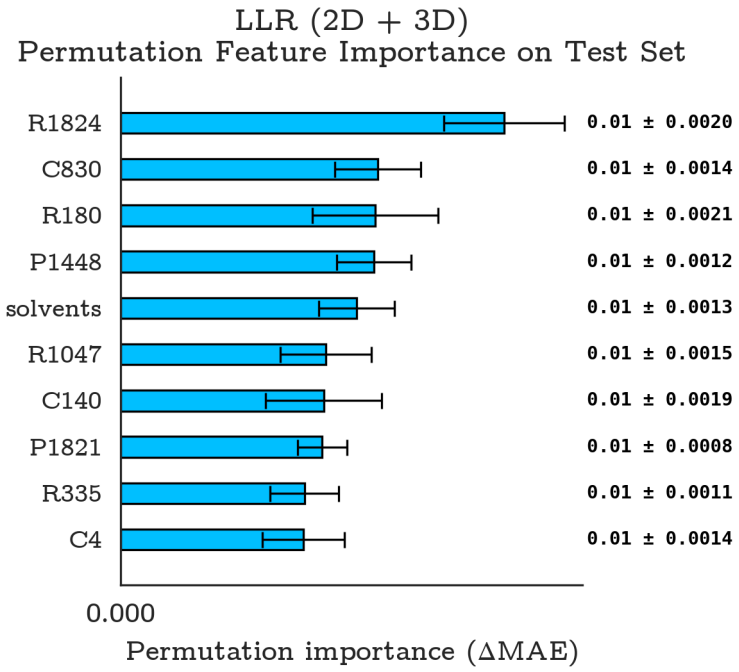


Figure 7.33: Permutation feature importance for the LLR(2D + 3D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted and error bars denote the SD.

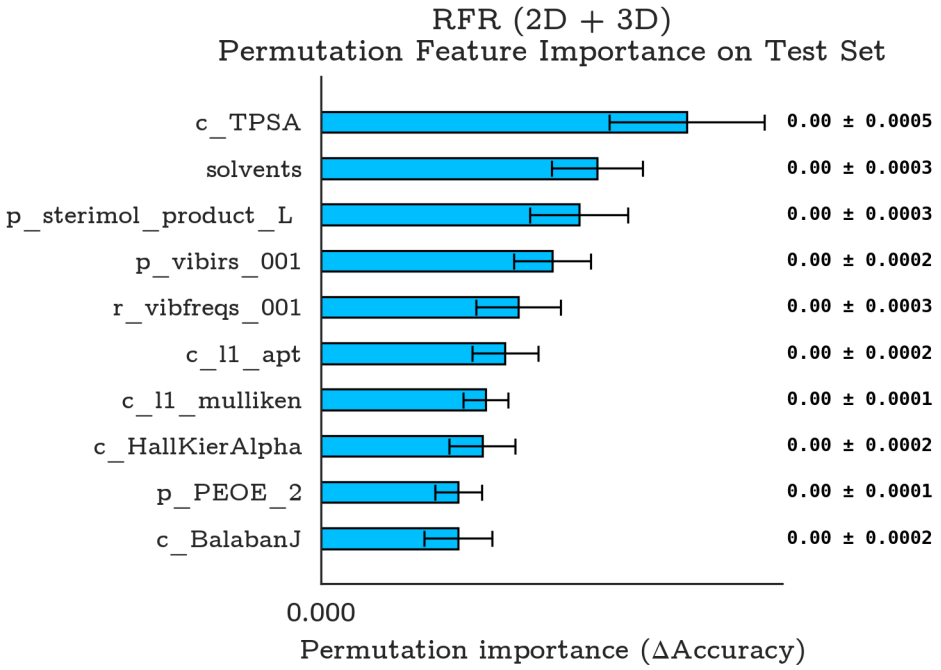


Figure 7.34: Permutation feature importance for the RFR(2D + 3D) model. Bars show the mean increase in test MAE (ΔMAE) when each feature's values are randomly permuted, and error bars denote the SD.

Table 7.5: Train and test set accuracies for all classification models across different sets of input features.

Model	Train accuracy (%)	Test accuracy (%)
RFC (2D)	82.2 ± 0.8	78.2 ± 1.8
ETC (2D)	88.3 ± 0.05	80.3 ± 1.9
GBC (2D)	86.8 ± 1.0	80.4 ± 2.6
RFC (3D)	83.5 ± 0.1	79.0 ± 10
ETC (3D)	83.7 ± 0.9	76.7 ± 1.2
GBC (3D)	89.1 ± 0.1	81.4 ± 1.1
RFC (2D + 3D)	82.9 ± 1.0	76.0 ± 1.9
ETC (2D + 3D)	85.7 ± 0.37	78.8 ± 1.8
GBC (2D + 3D)	84.0 ± 1.8	79.4. ± 1.2

Table 7.6: Optimum set of hyperparameters found for classification models with 2D input features.

Model	Hyperparameters	Values
RFC	Max_features	0.4
	Min_samples_leaf	15
	Min_samples_split	50
	Max_depth	12
	n_estimators	800
GBC	Learning_rate	0.1
	Max_depth	5
	Max_features	0.9
	Min_samples_leaf	60
	Min_samples_split	40
	n_estimators	800
ETC	Max_features	0.6
	Min_samples_leaf	8
	Min_samples_split	11
	Max_depth	20
	n_estimators	1000

Table 7.7: Optimum set of hyperparameters for classification models with 3D input features.

Model	Hyperparameters	Values
RFC	Max_features	0.8
	Min_samples_split	11
	Max_depth	5
	n_estimators	1200
GBC	Learning_rate	0.05
	Max_depth	2
	Max_features	0.7
	Min_samples_leaf	30
	n_estimators	400
ETC	Max_features	0.6
	Min_samples_leaf	13
	Min_samples_split	10
	Max_depth	12
	n_estimators	800

Table 7.8: Optimum set of hyperparameters for classification models with 2D + 3D input features.

Model	Hyperparameters	Values
RFC	Max_features	0.4
	Min_samples_leaf	15
	Min_samples_split	50
	Max_depth	12
	n_estimators	600
GBC	Learning_rate	0.01
	Max_depth	2
	Max_features	0.3
	Min_samples_leaf	60
	Min_samples_split	80
	n_estimators	2000
ETC	Max_features	0.6
	Min_samples_leaf	20
	Min_samples_split	11
	Max_depth	21
	n_estimators	1000

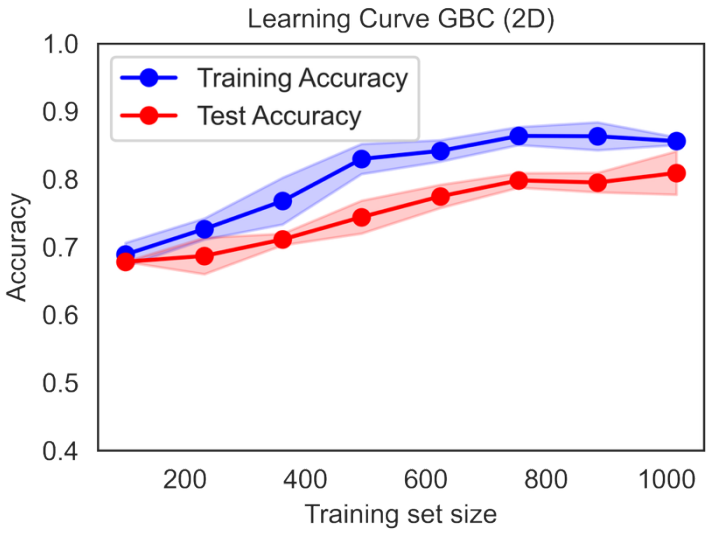


Figure 7.35: Learning curves for the GBC(2D) model. Blue markers show the training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

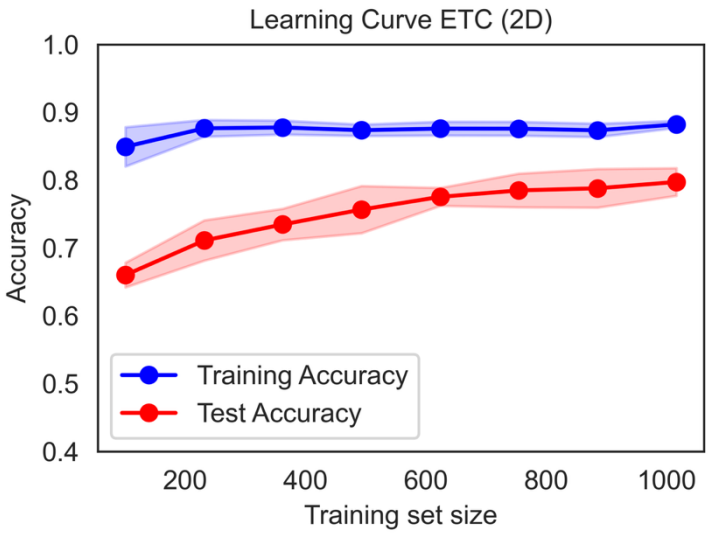


Figure 7.36: Learning curves for the ETC(2D) model. Blue markers show the training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

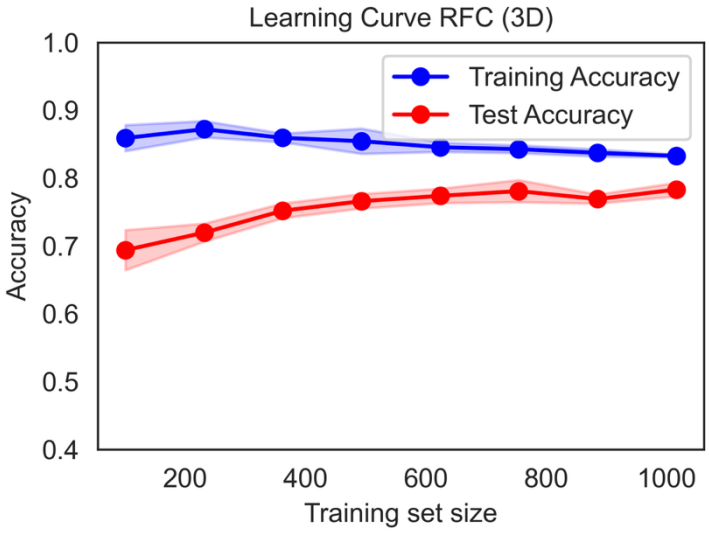


Figure 7.37: Learning curves for the RFC(3D) model. Blue markers show the training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

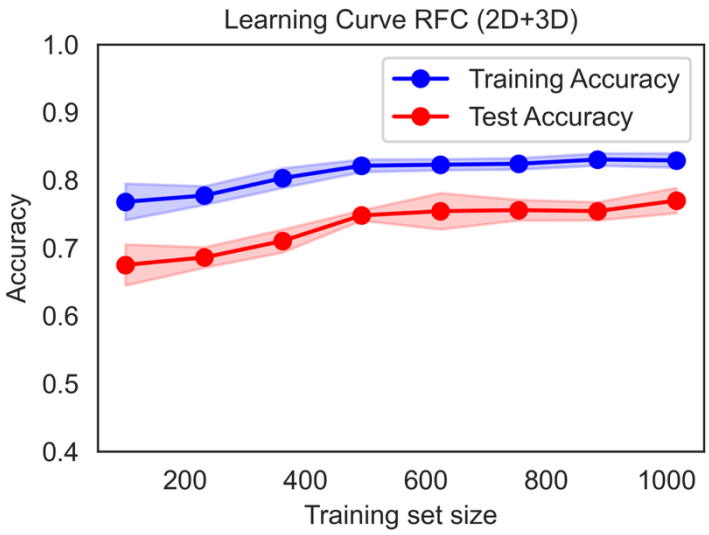


Figure 7.38: Learning curves for the RFC(2D + 3D) model. Blue markers show the training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

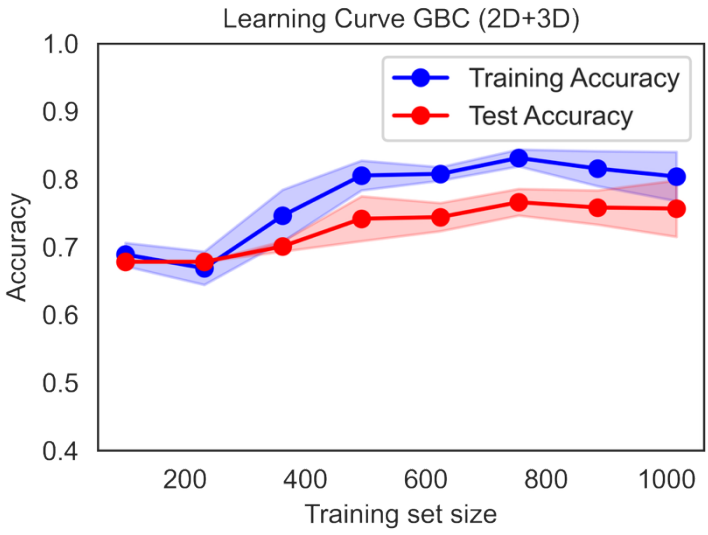


Figure 7.39: Learning curves for the GBC(2D + 3D) model. Blue markers show the training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

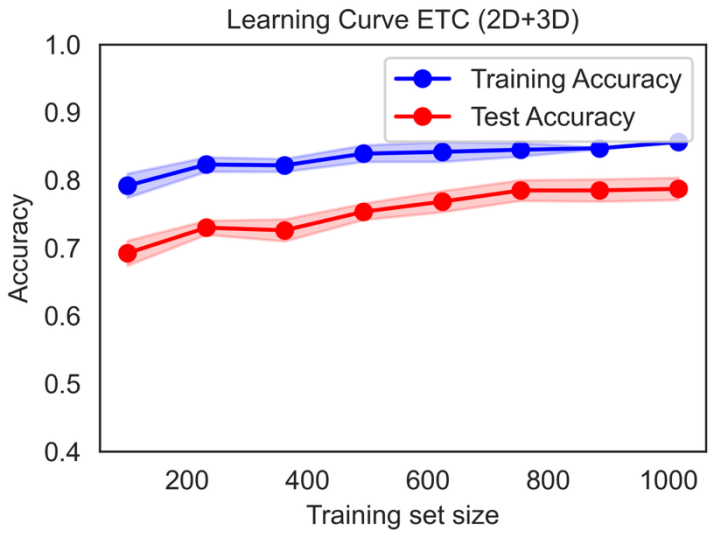


Figure 7.40: Learning curves for the ETC(2D + 3D) model. Blue markers show the training accuracy and the red markers show the CV test accuracy. The shaded bands denote the SD.

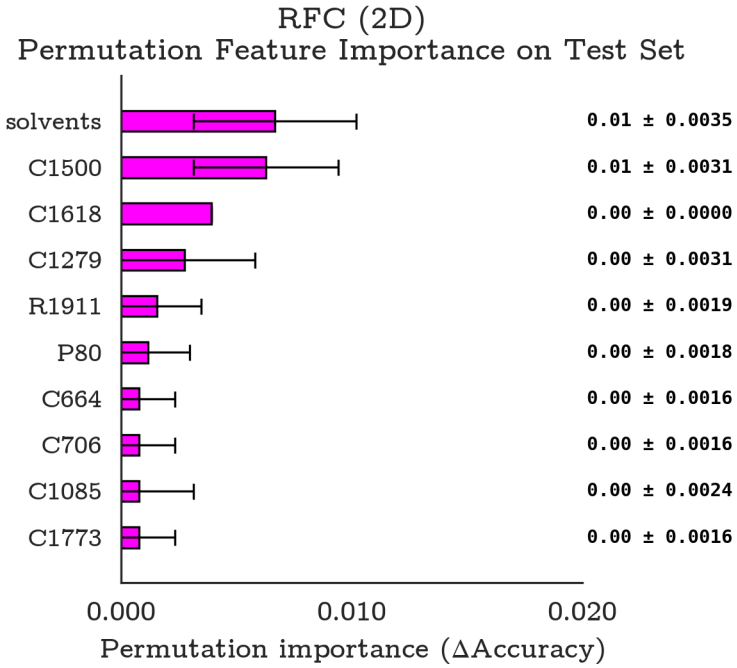


Figure 7.41: Permutation feature importance for the RFC(2D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

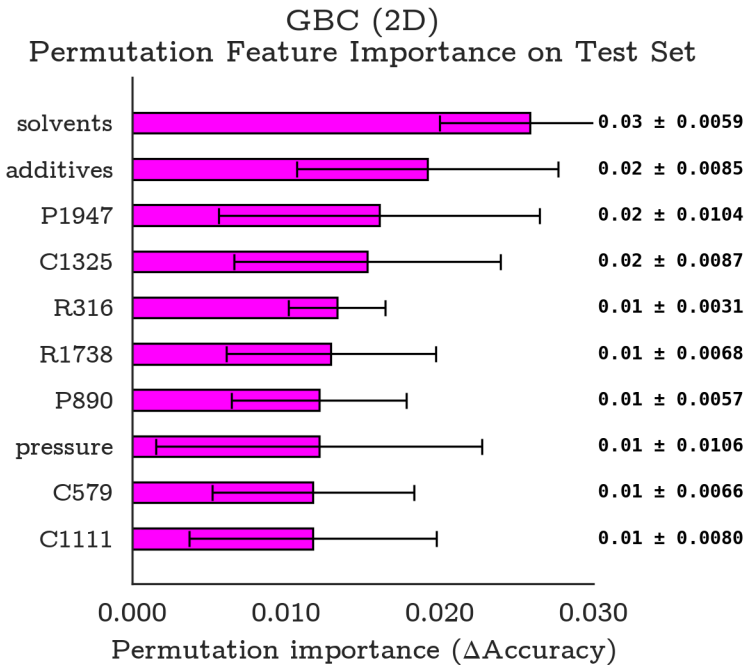


Figure 7.42: Permutation feature importance for the GBC(2D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

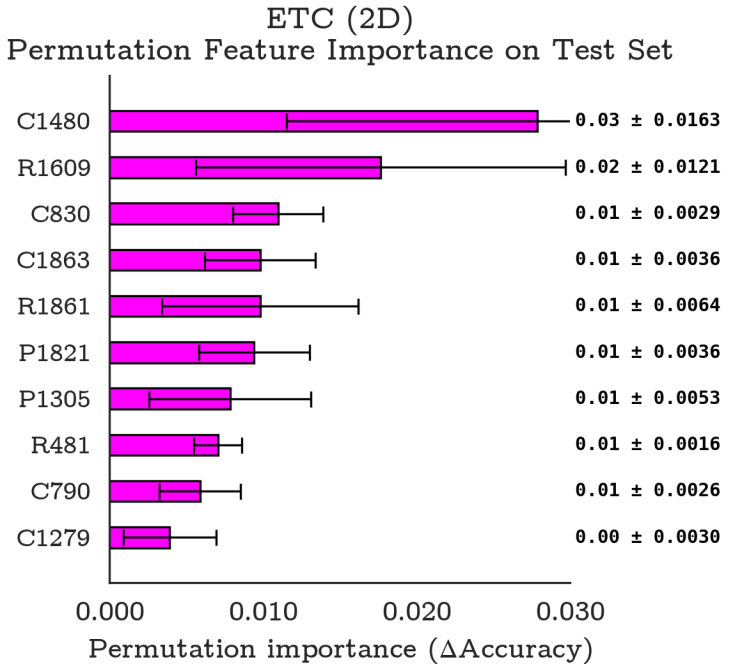


Figure 7.43: Permutation feature importance for the ETC(2D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

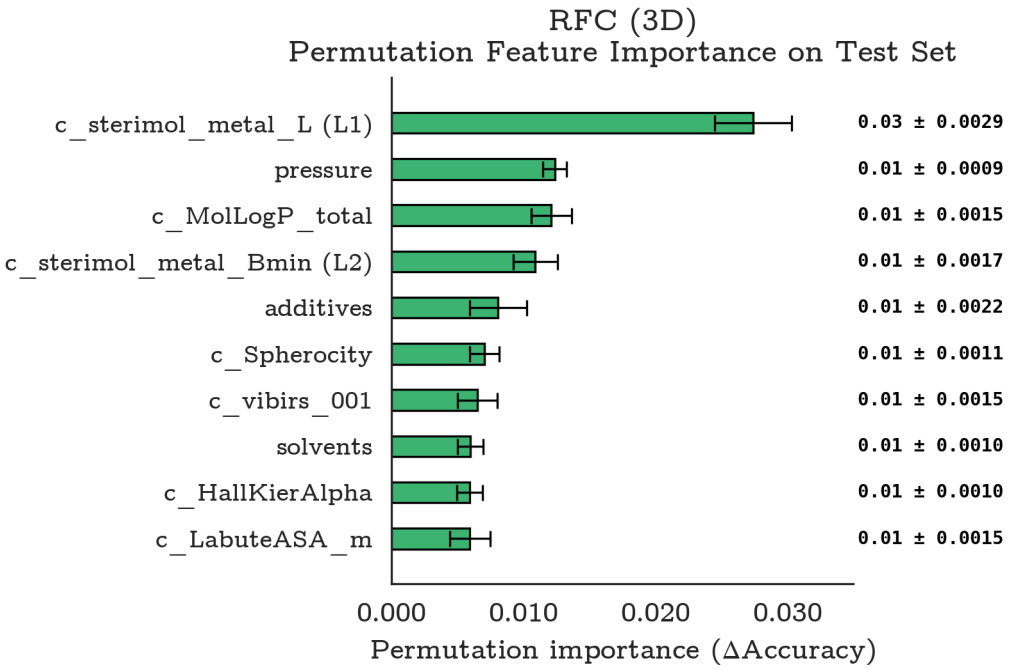


Figure 7.44: Permutation feature importance for the RFC(3D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

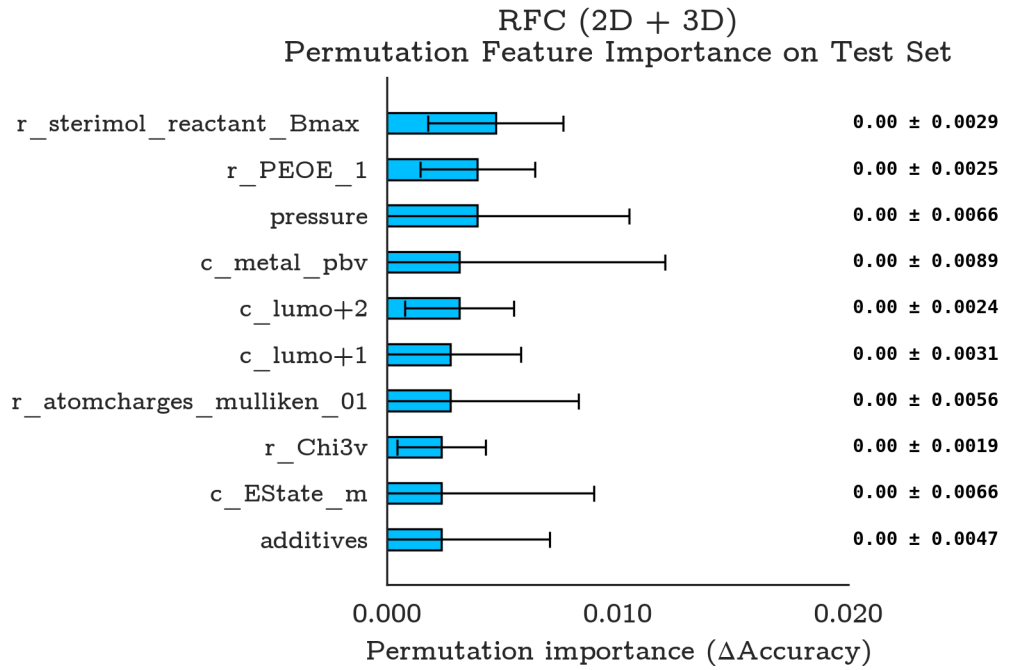


Figure 7.45: Permutation feature importance for the RFC(2D + 3D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

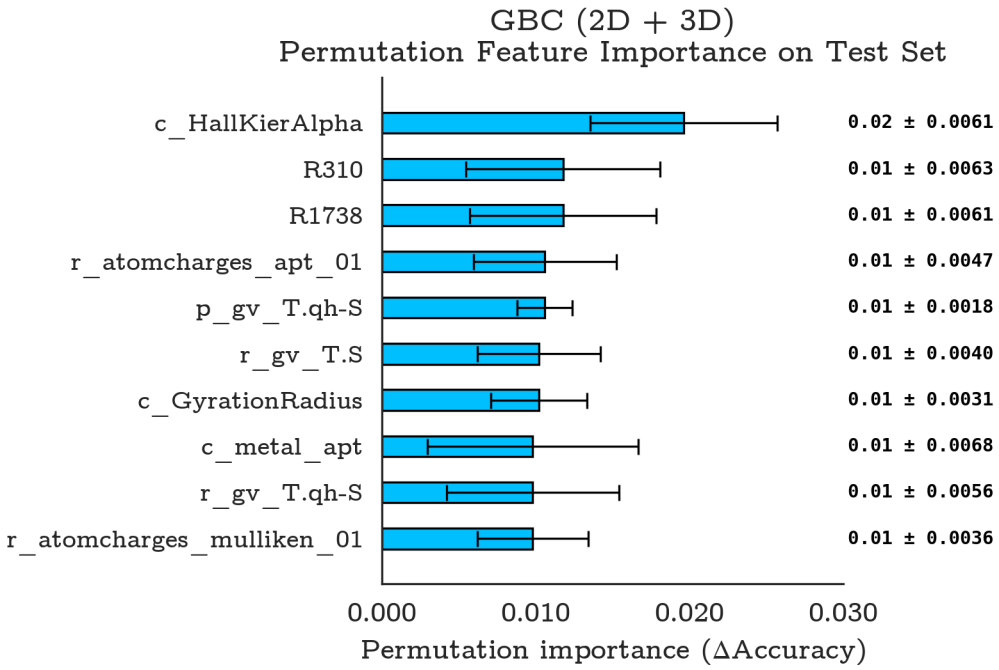


Figure 7.46: Permutation feature importance for the GBC(2D + 3D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

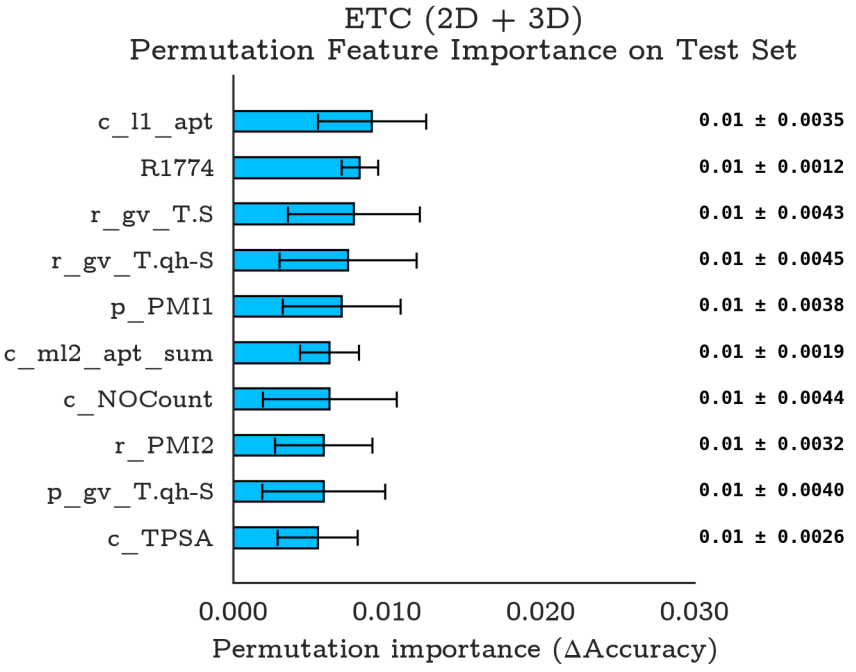


Figure 7.47: Permutation feature importance for the ETC(2D + 3D) model. Bars show the mean increase in test accuracy (Δ accuracy) when each feature's values are randomly permuted and error bars denote the SD.

Table 7.9: Train and test set accuracies for classification models with 3D input features with additional SMOTE data added to the training set.

Model	Train accuracy (%)	Test accuracy (%)
ETC SMOTE (3D)	84.9 ± 1.15	78.6 ± 1.6
GBC SMOTE (3D)	90.3 ± 0.26	80.7 ± 1.5

Table 7.10: Optimum set of hyperparameters for classification models 3D input features and additional SMOTE data added to the training set.

Model	Hyperparameters	Values
GBC	Learning_rate	0.007
	Max_depth	2
	Max_features	0.9
	Min_samples_leaf	40
	n_estimators	500
ETC	Max_features	0.85
	Min_samples_leaf	10
	Min_samples_split	11
	Max_depth	8
	n_estimators	800

1.1 CHAPTER 3 APPENDIX

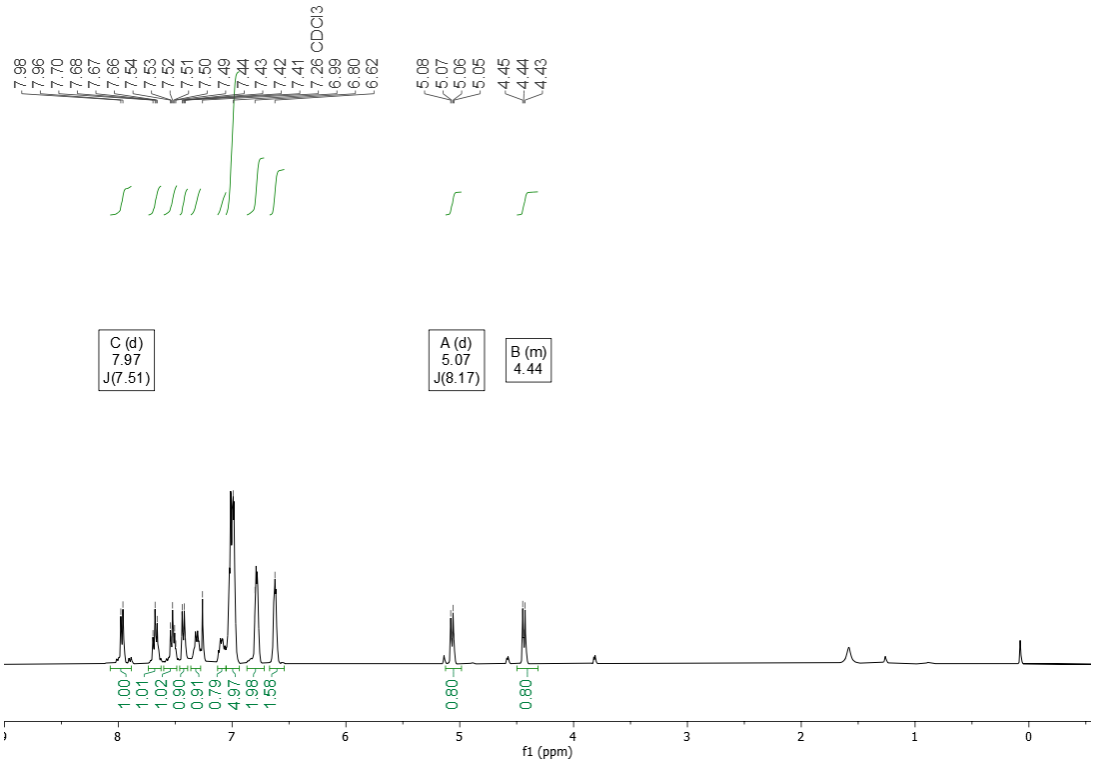


Figure 7.48: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of (*S*,*S*) (*R*,*R*) 2,3-diphenyl-indan-1-one (**P3.1**).

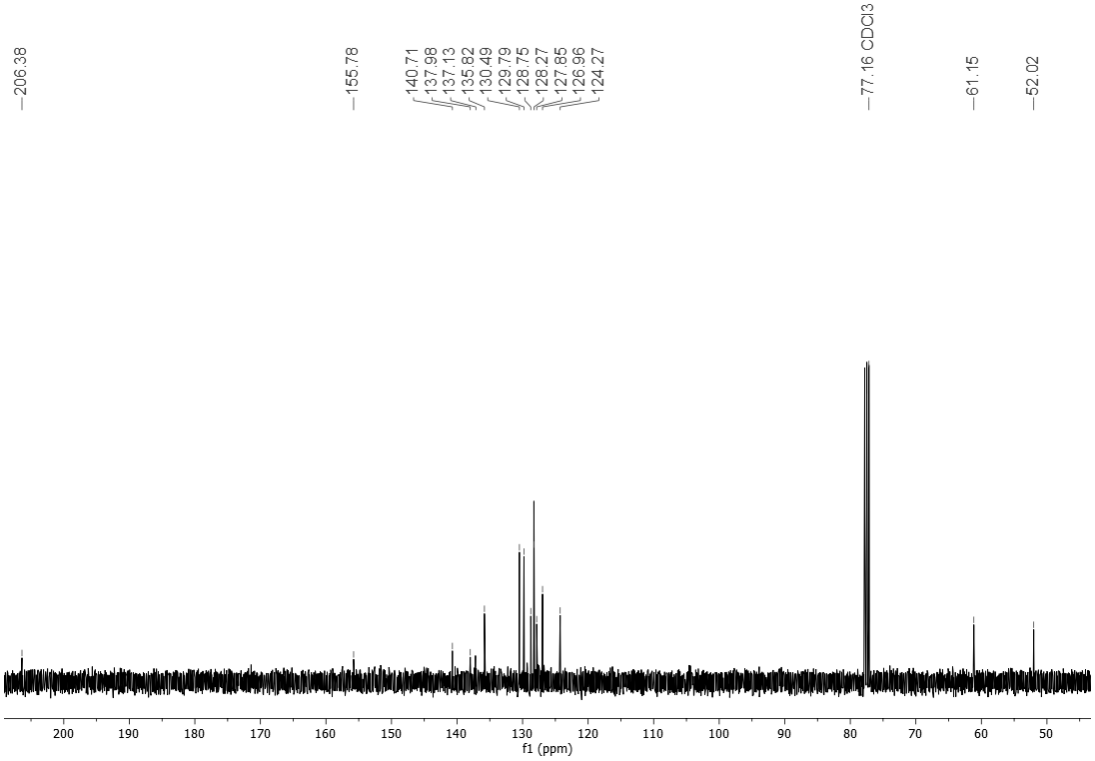


Figure 7.49: $^{13}C\{^{1}H\}$ NMR (101 MHZ, $CDCl_3$) spectrum of (*S*,*S*) (*R*,*R*) 2,3-diphenyl-indan-1-one (**P3.1**).

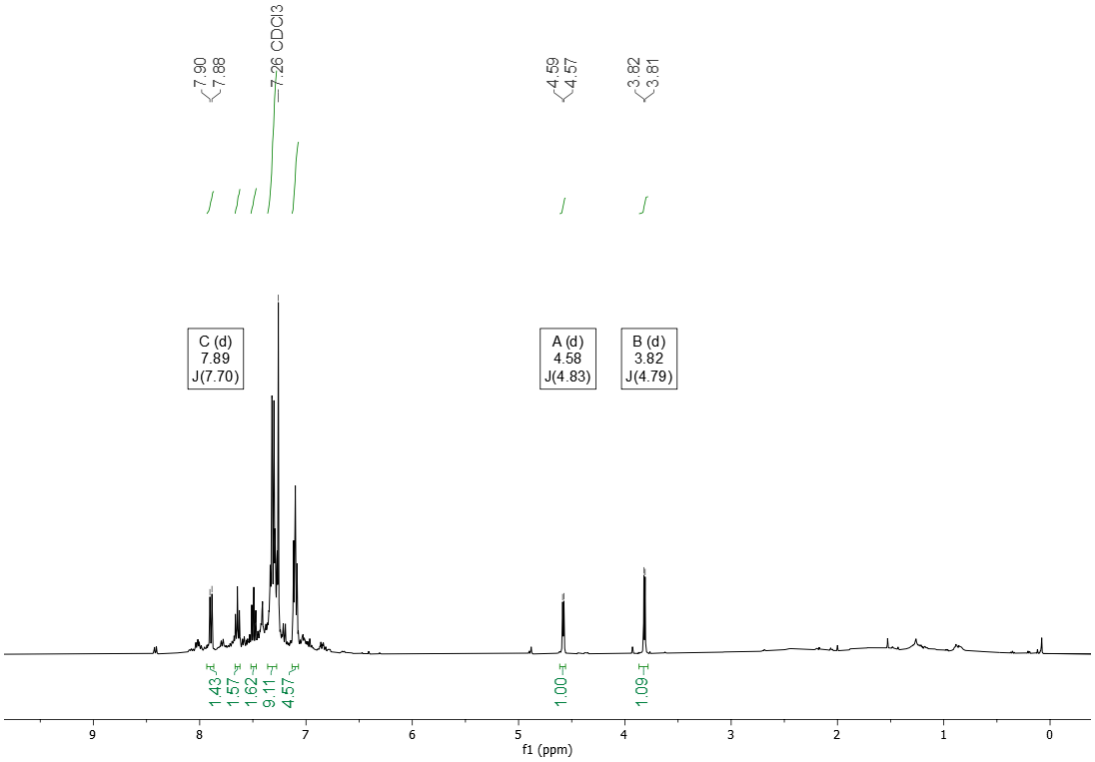


Figure 7.50: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of (*S*,*R*) (*R*,*S*) 2,3-diphenyl-indan-1-one.

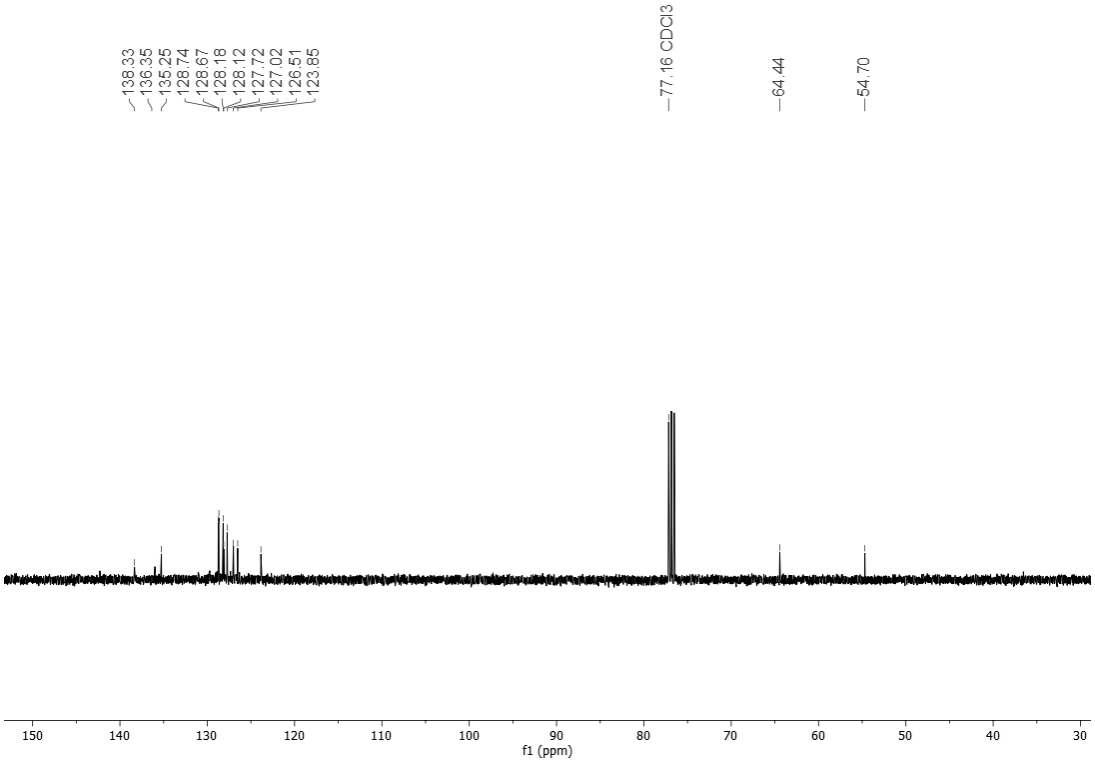


Figure 7.51: $^{13}C\{^{1}H\}$ NMR (101 MHZ, $CDCl_3$) spectrum of (*S*,*R*) (*R*,*S*) 2,3-diphenyl-indan-1-one.

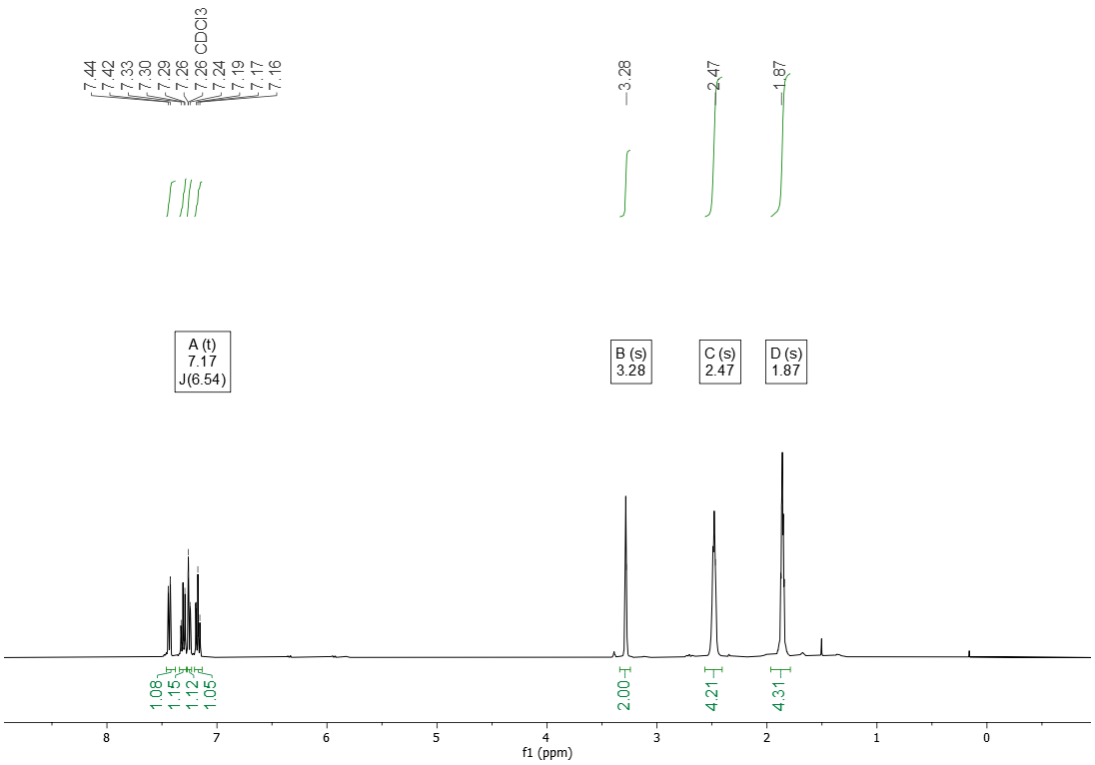


Figure 7.52: : ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of 1,2,3,4-tetrahydro-9H-fluorene (**S3.2**).

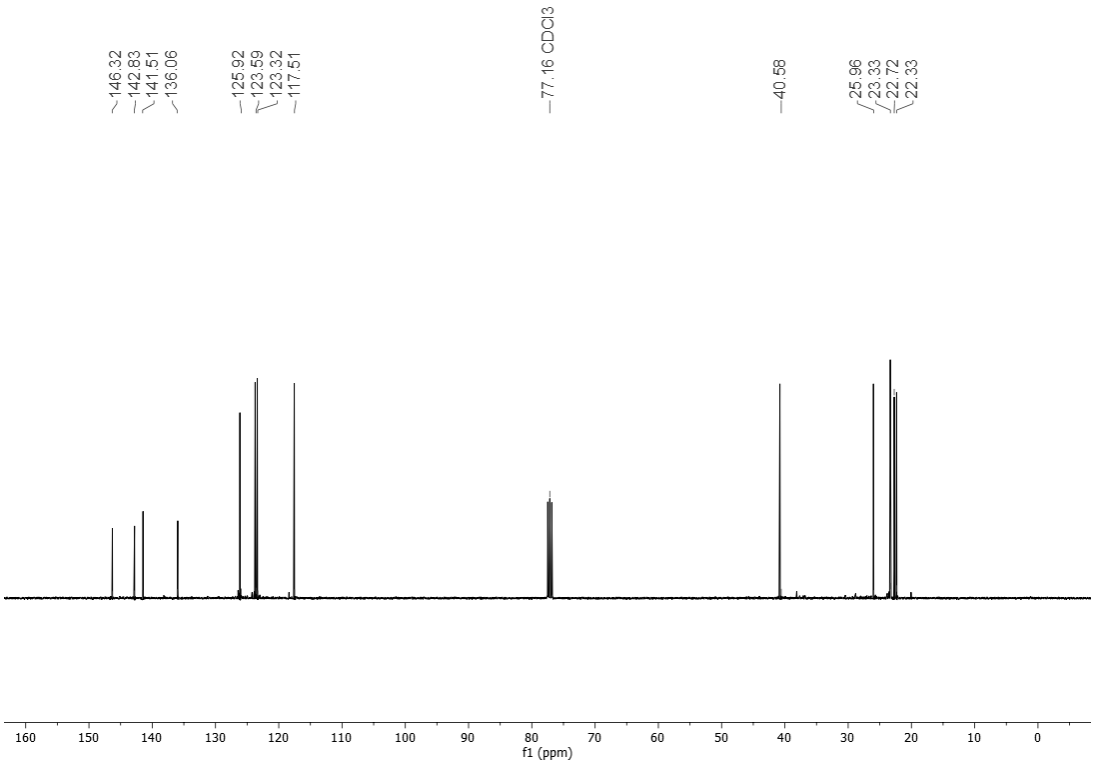


Figure 7.53: $^{13}C\{^{1}H\}$ NMR (101 MHZ, $CDCl_3$) spectrum of 1,2,3,4-tetrahydro-9H-fluorene (**S3.2**)

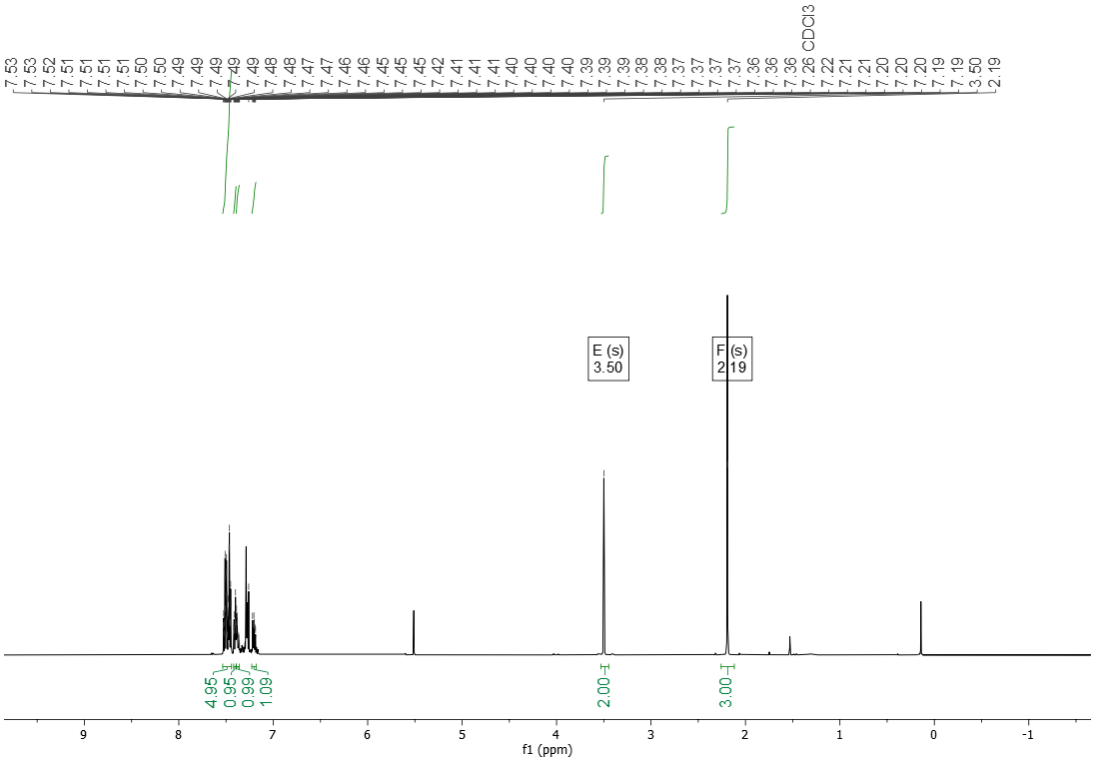


Figure 7.54: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of 2-Methyl-3-phenyl-1H-indene (**S3.3**).

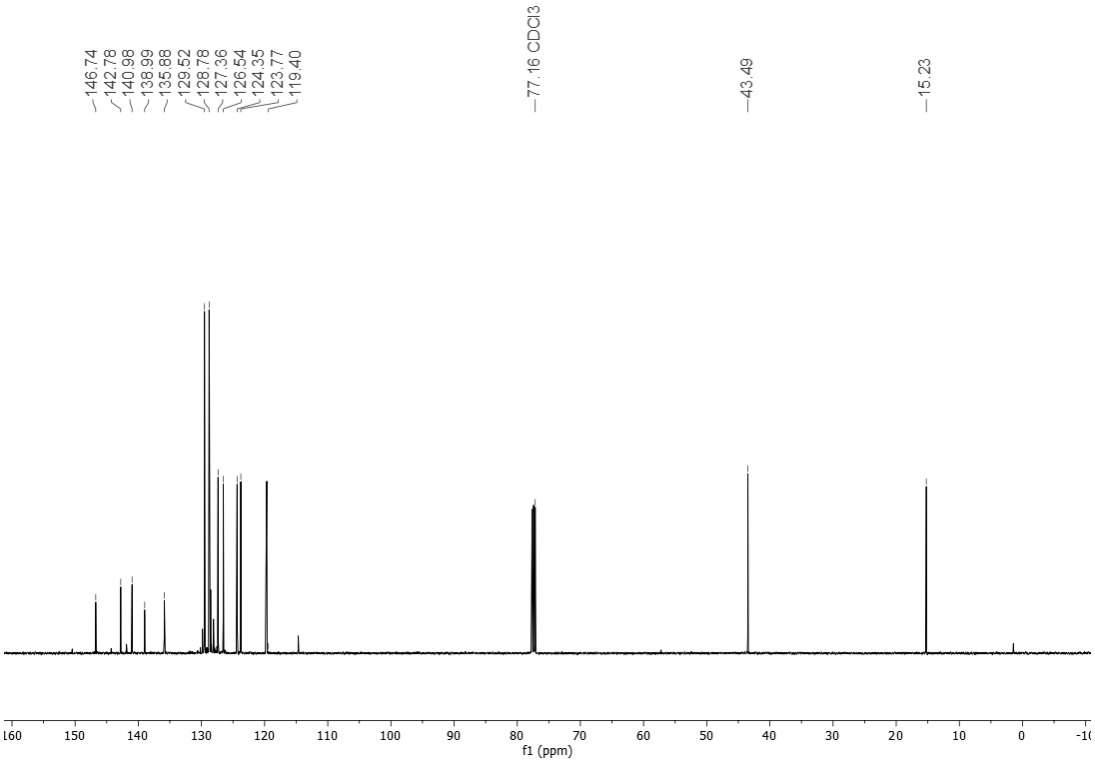


Figure 7.55: : $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum of 2-Methyl-3-phenyl-1H-indene (**S3.3**).

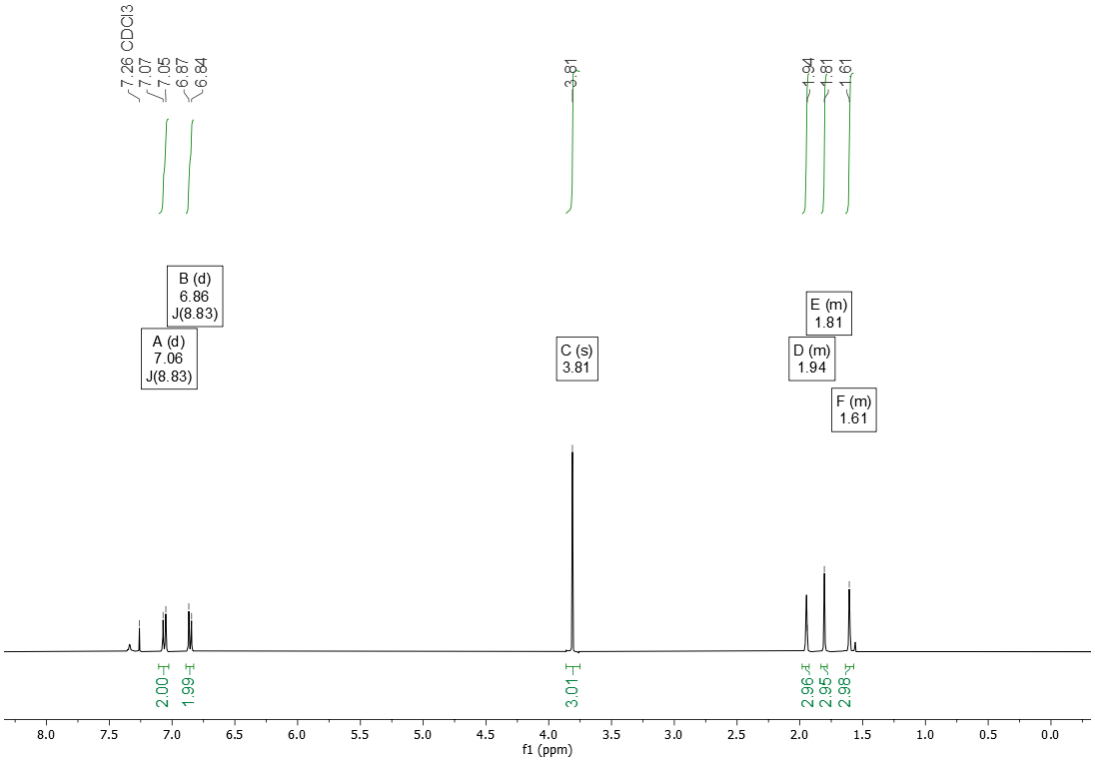


Figure 7.56: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of 1-Methoxy-4-(3-methylbut-2-en-2-yl)benzene (**S3.4**).

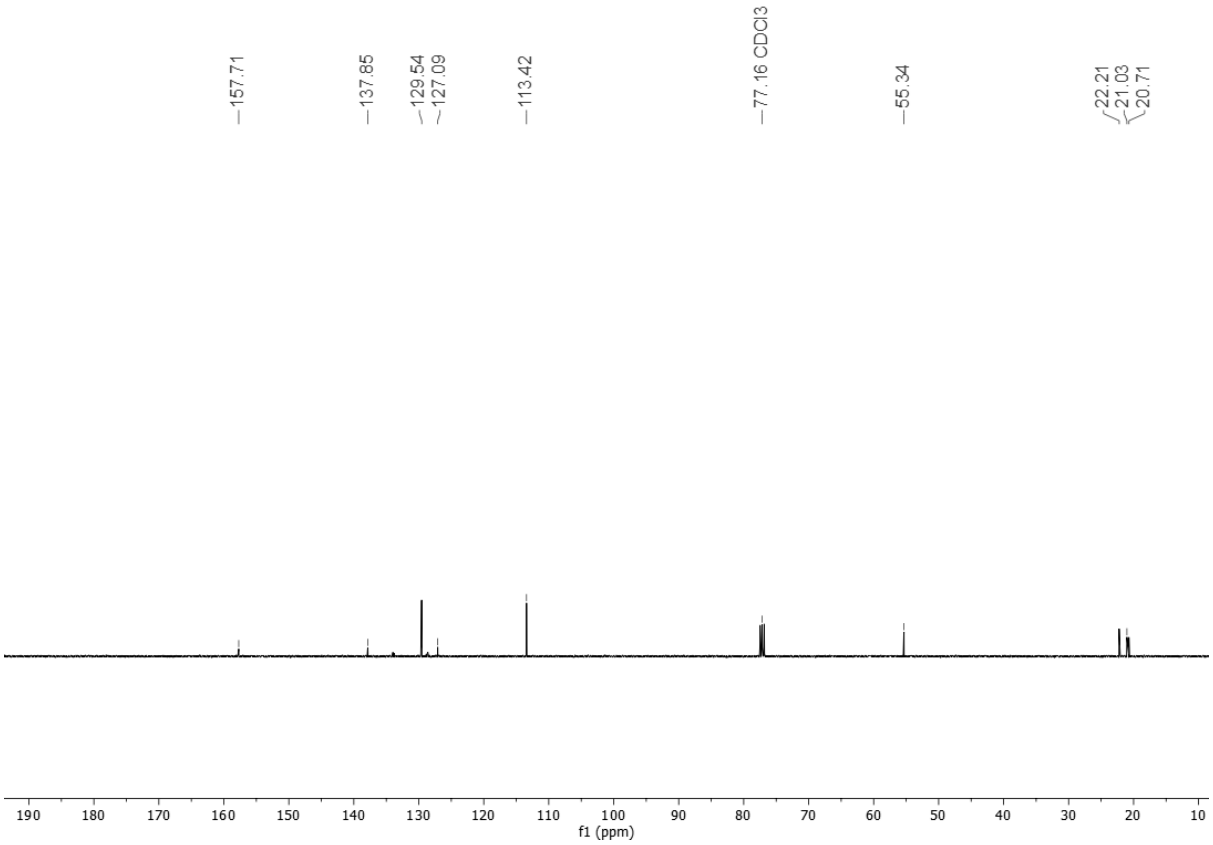


Figure 7.57: $^{13}C\{^{1}H\}$ NMR (101 MHZ, $CDCl_3$) spectrum of 1-Methoxy-4-(3-methylbut-2-en-2-yl)benzene (**S3.4**).

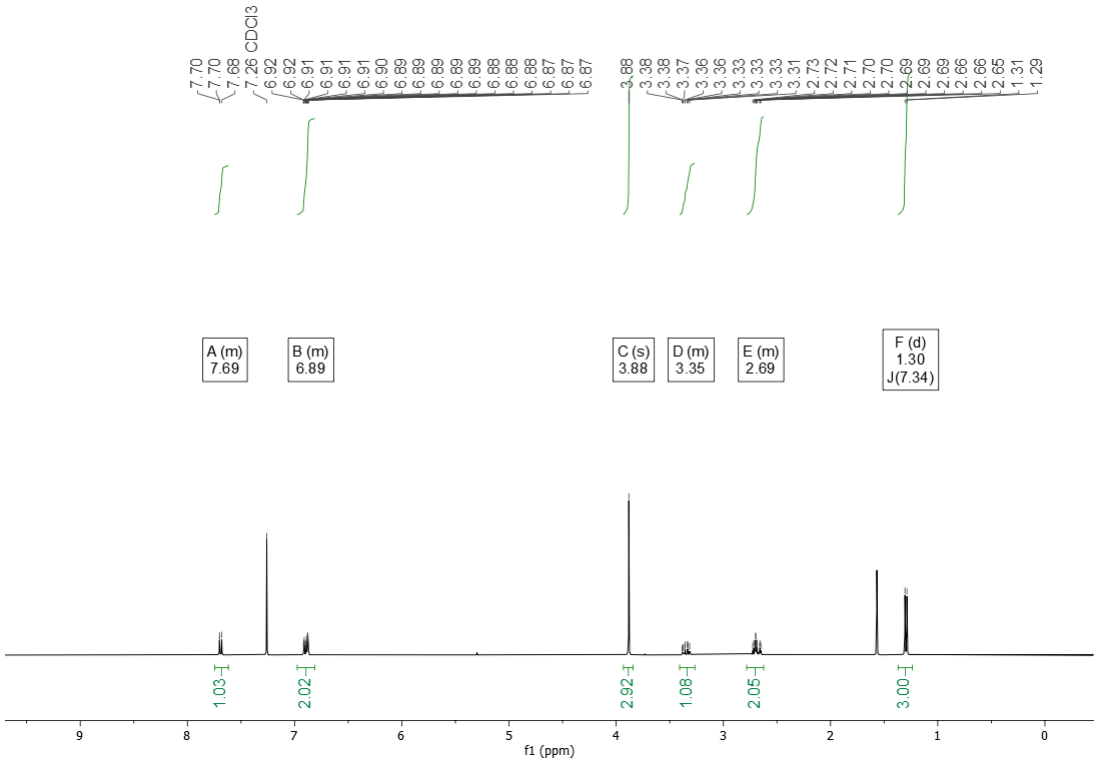


Figure 7.58: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of 5-methoxy-2-methyl-2,3-dihydro-1*H*-indenone.

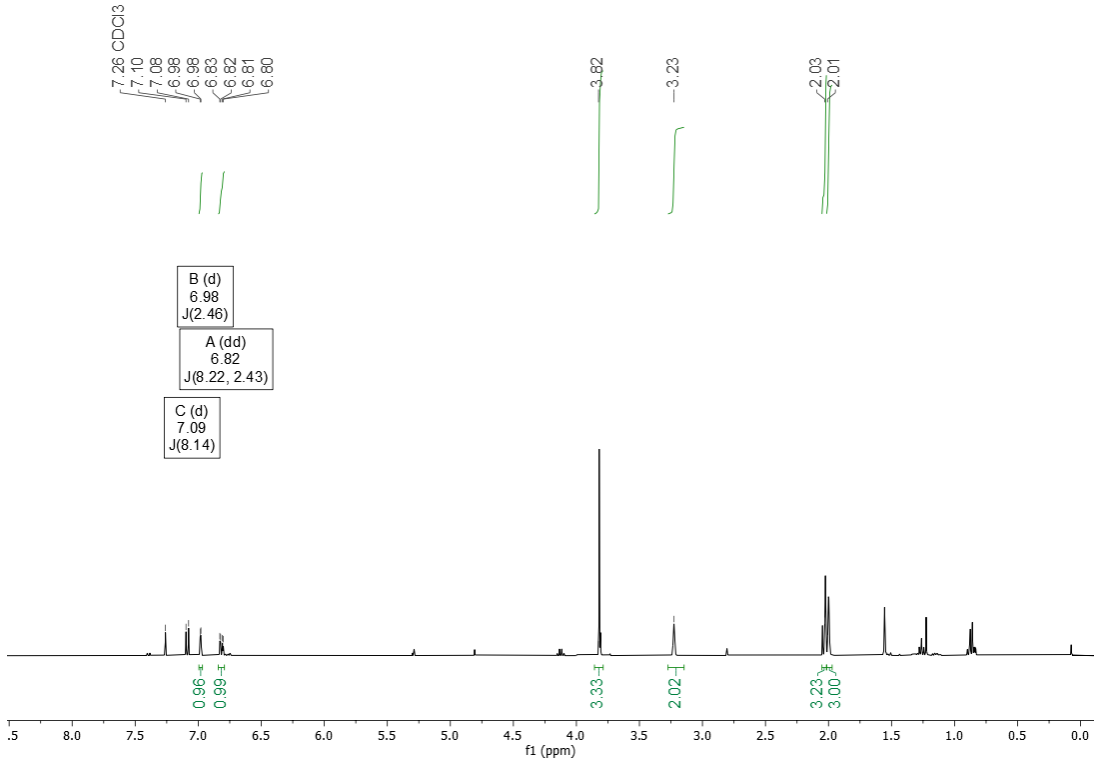


Figure 7.59: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of 6-Methoxy-2,3-dimethyl-1*H*-indene (**S3.5**).

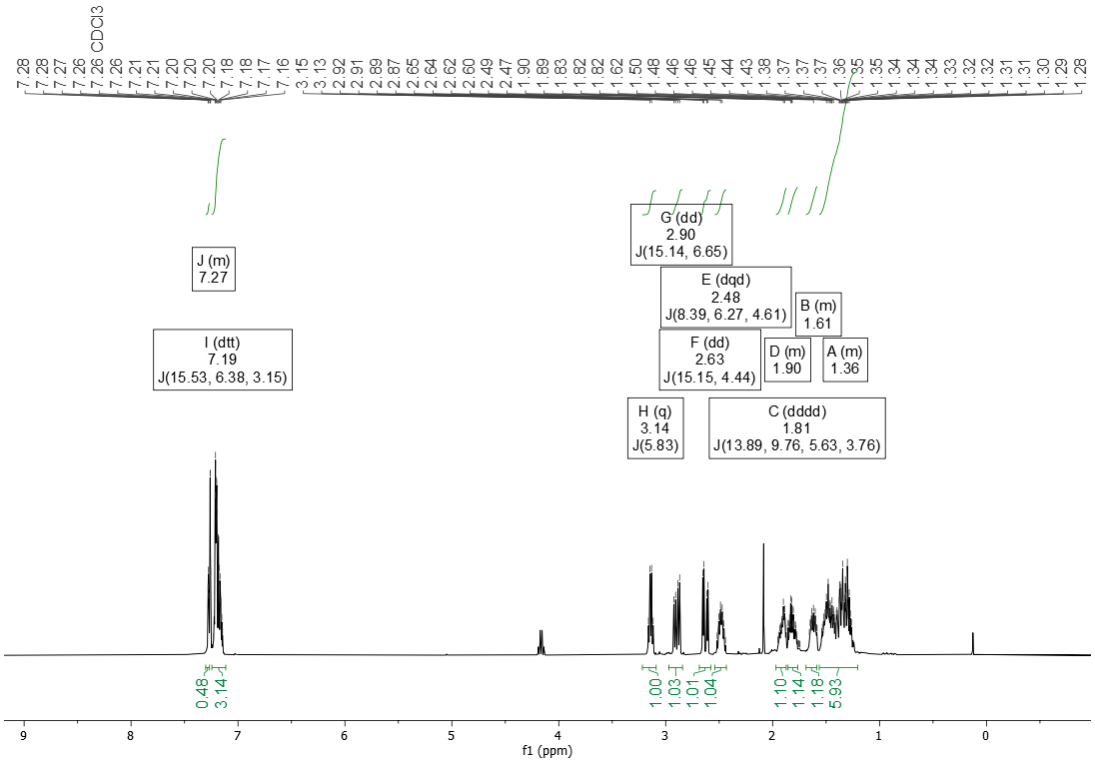


Figure 7.60: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of 1 cis-1,2,3,4,4a,9,9a-heptahydro-4aH-flourene (**P3.2**).

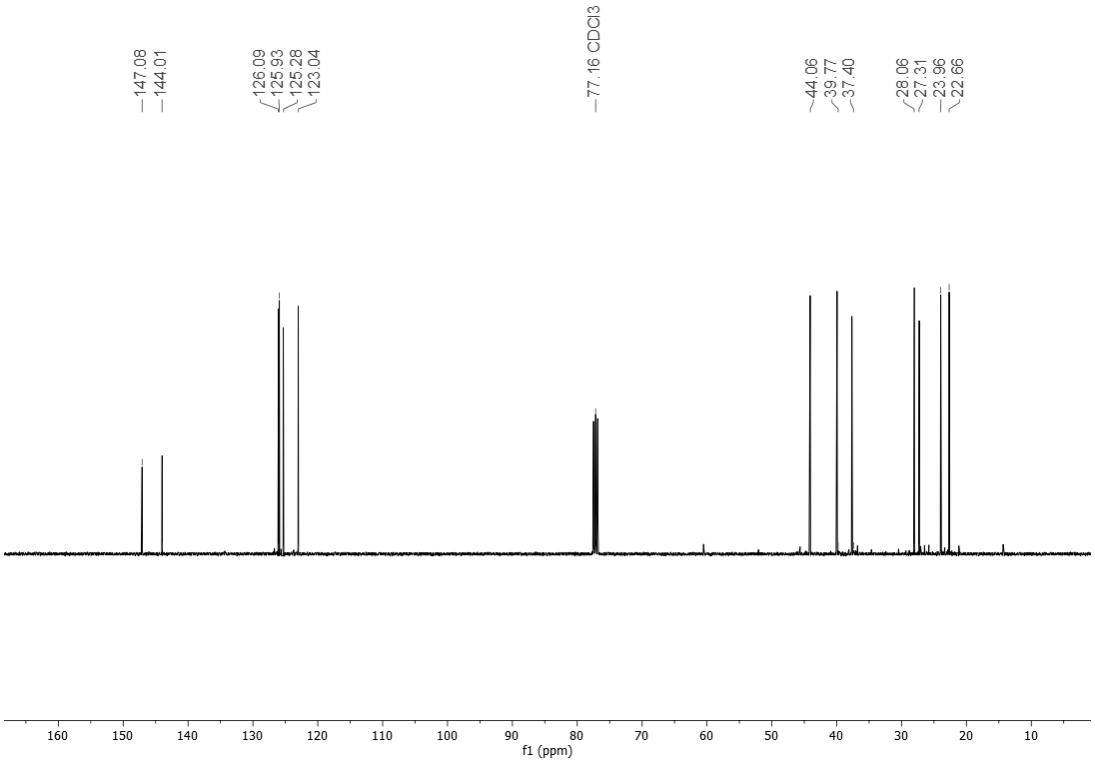


Figure 7.61: $^{13}C\{^{1}H\}$ NMR (101 MHZ, $CDCl_3$) spectrum of 1 cis-1,2,3,4,4a,9,9a-heptahydro-4aH-flourene (**P3.2**).

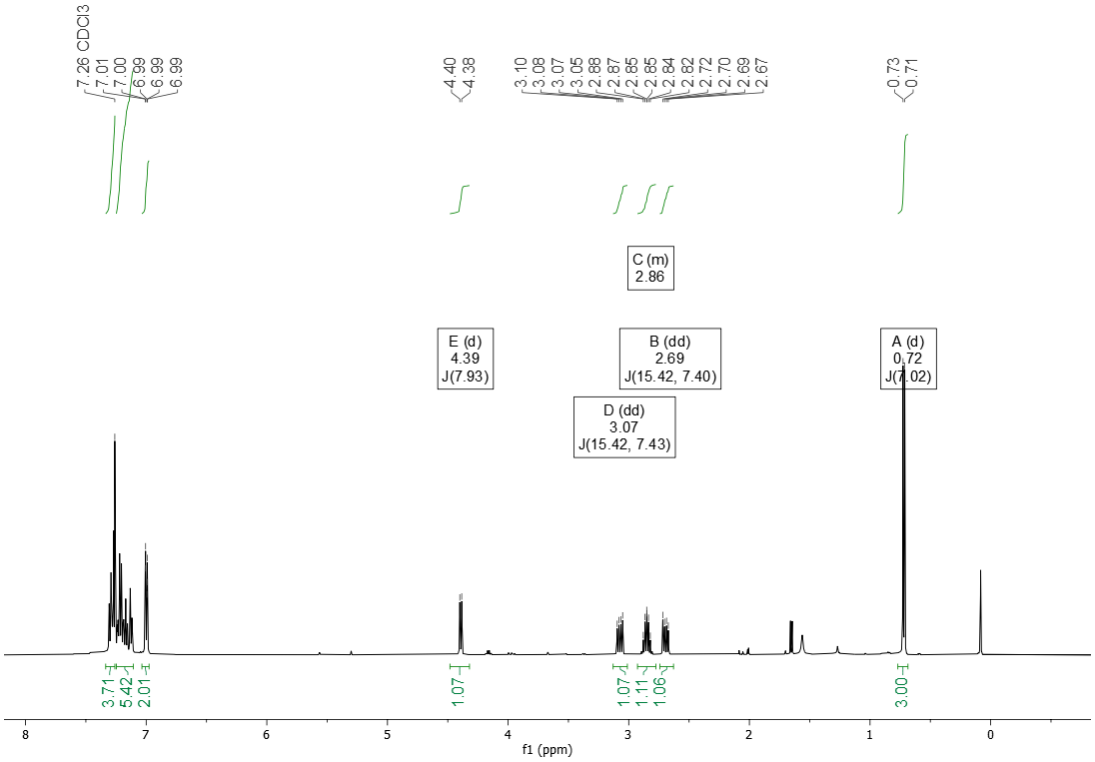


Figure 7.62: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of cis-1-phenyl-2-methylindane (**P3.3**).

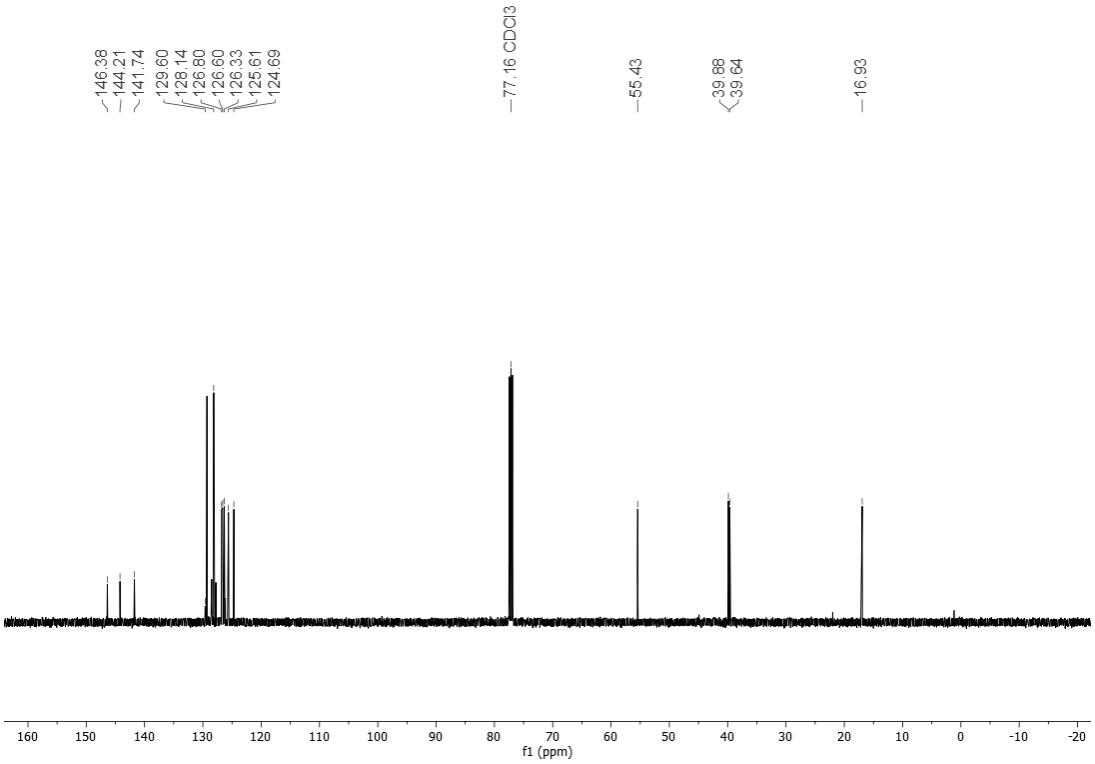


Figure 7.63: $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum of cis-1-phenyl-2-methylindane (**P3.3**).

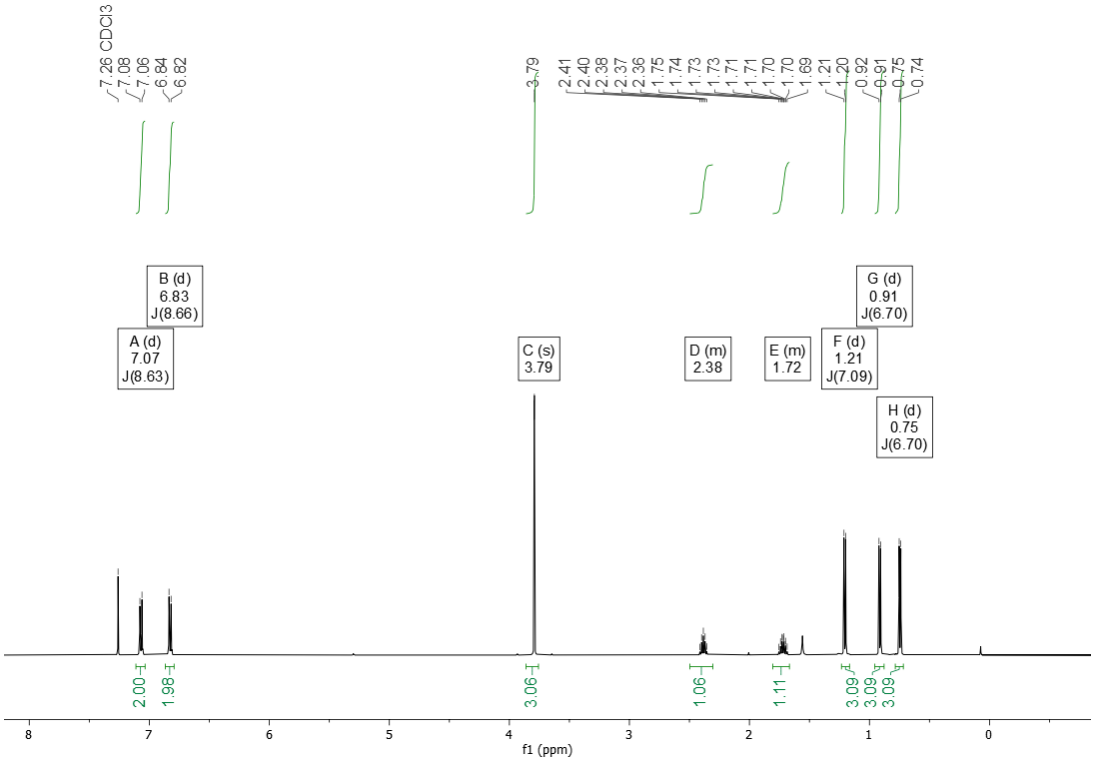


Figure 7.64: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of 1-Methoxy-4-(3-methylbutan-2-yl)benzene (**P3.4**).

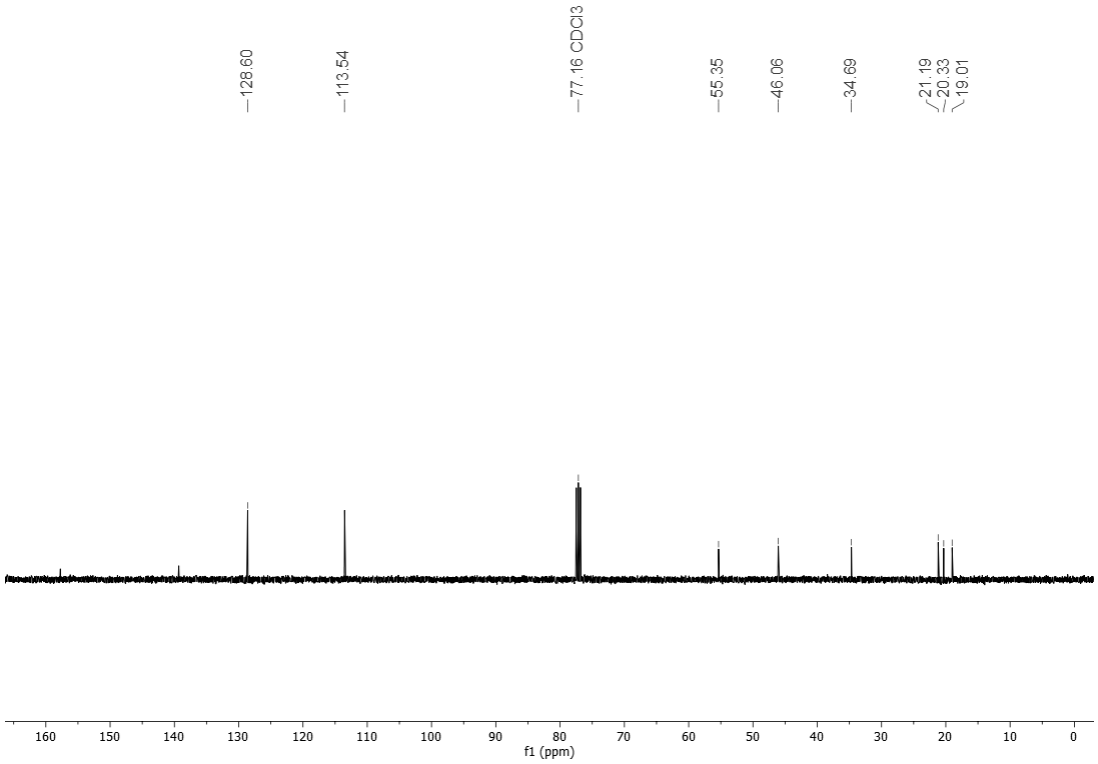


Figure 7.65: : $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum of 1-Methoxy-4-(3-methylbutan-2-yl)benzene (**P3.4**).

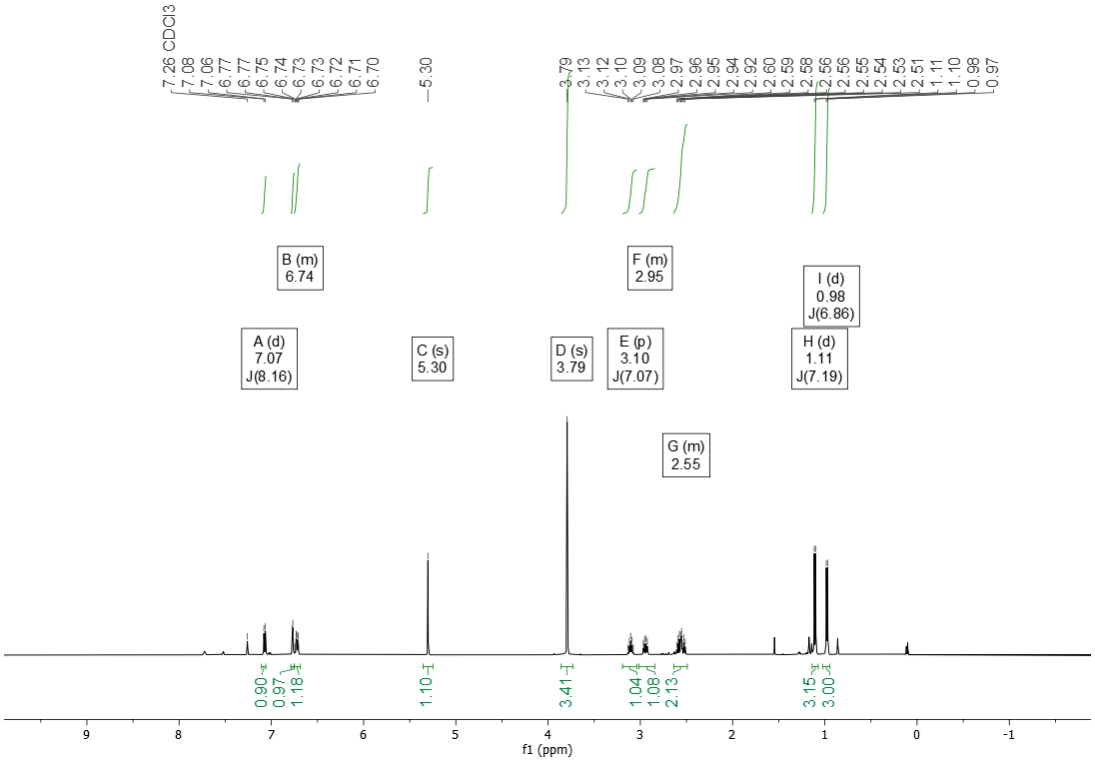


Figure 7.66: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of 5-Methoxy-1,2-dimethyl-2,3-dihydro-1H-indene (**P3.5**).

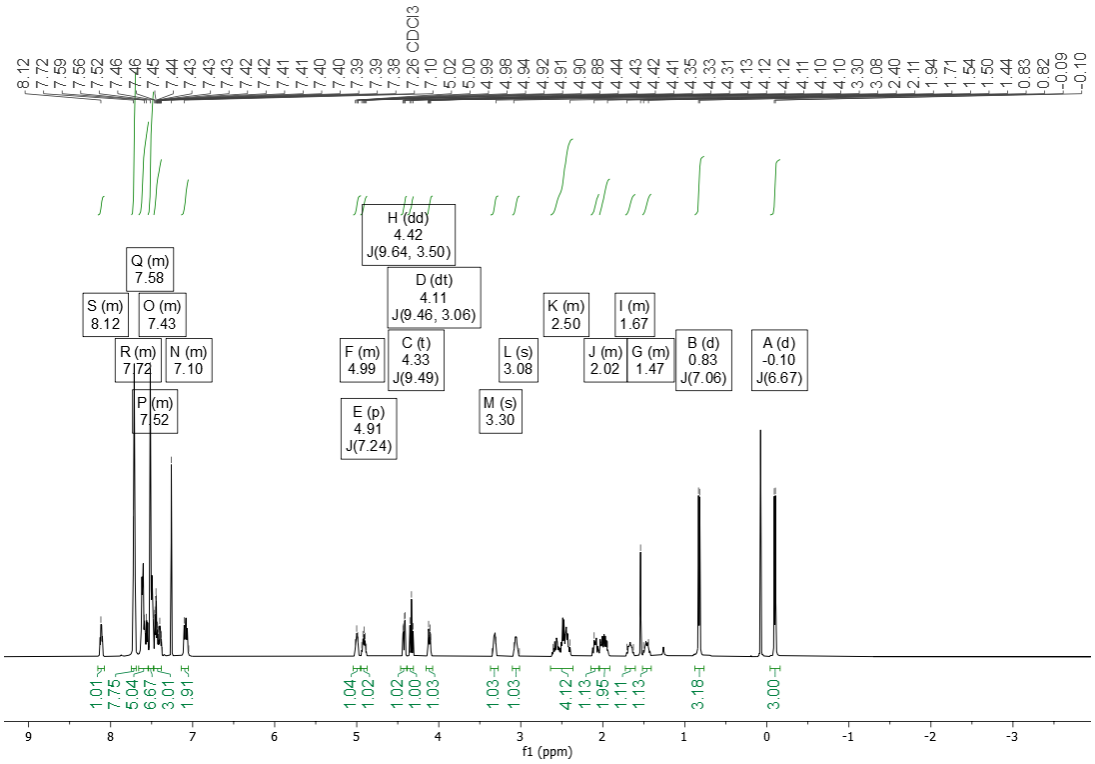


Figure 7.67: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of [Ir(**L3.1**)COD]BAr^{F}_{4}.

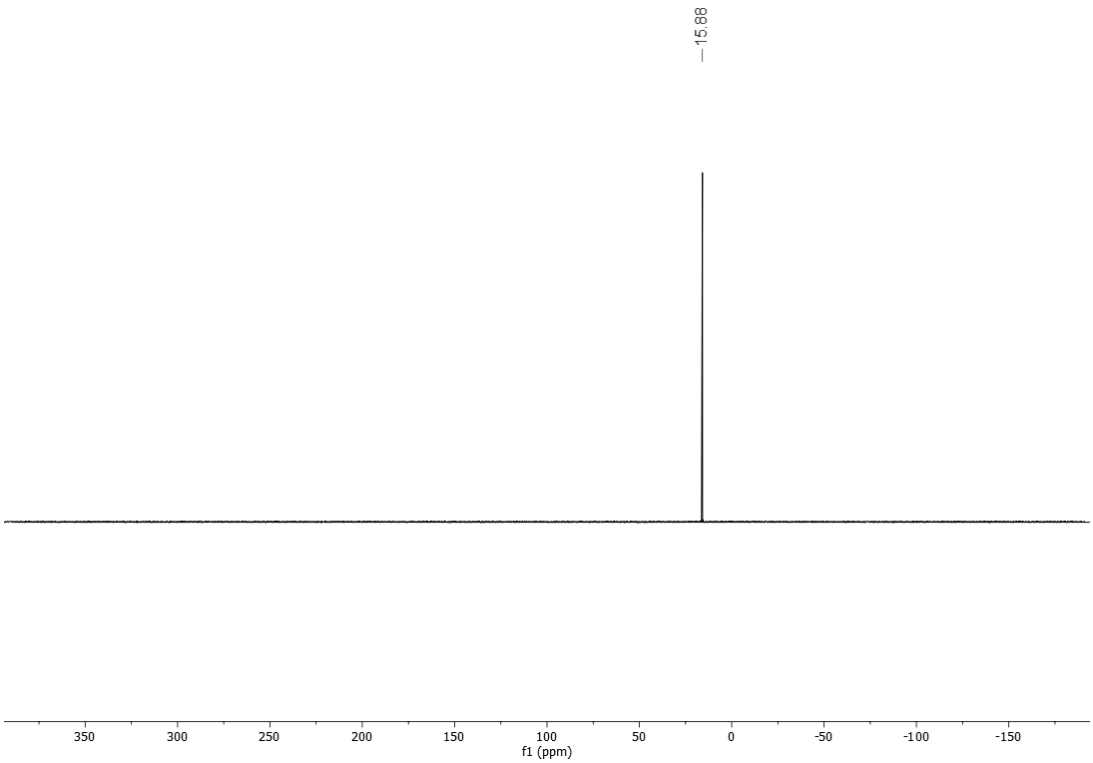


Figure 7.68: $^{31}P\{^{1}H\}$ NMR (203 MHZ, $CDCl_3$) spectrum of [Ir(**L3.1**)COD]BAr^F_4.

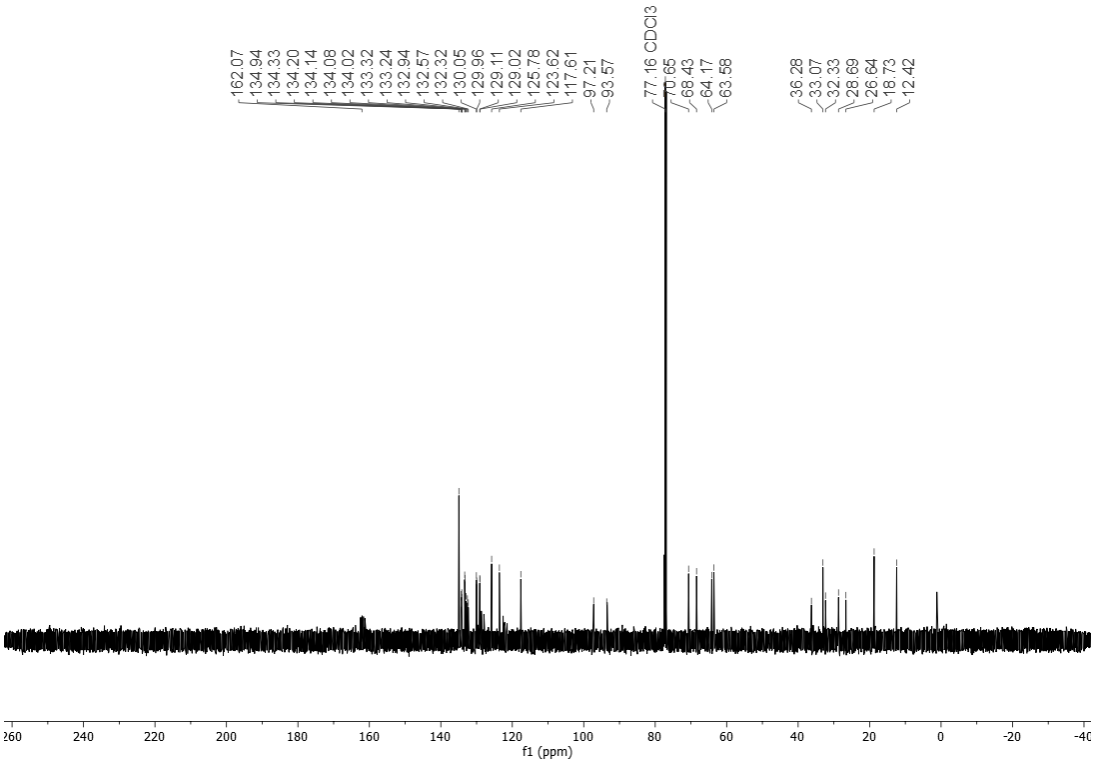


Figure 7.69: $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum of [Ir(**L3.1**)COD]BAr^F_4.

1.2 CHAPTER 4 APPENDIX

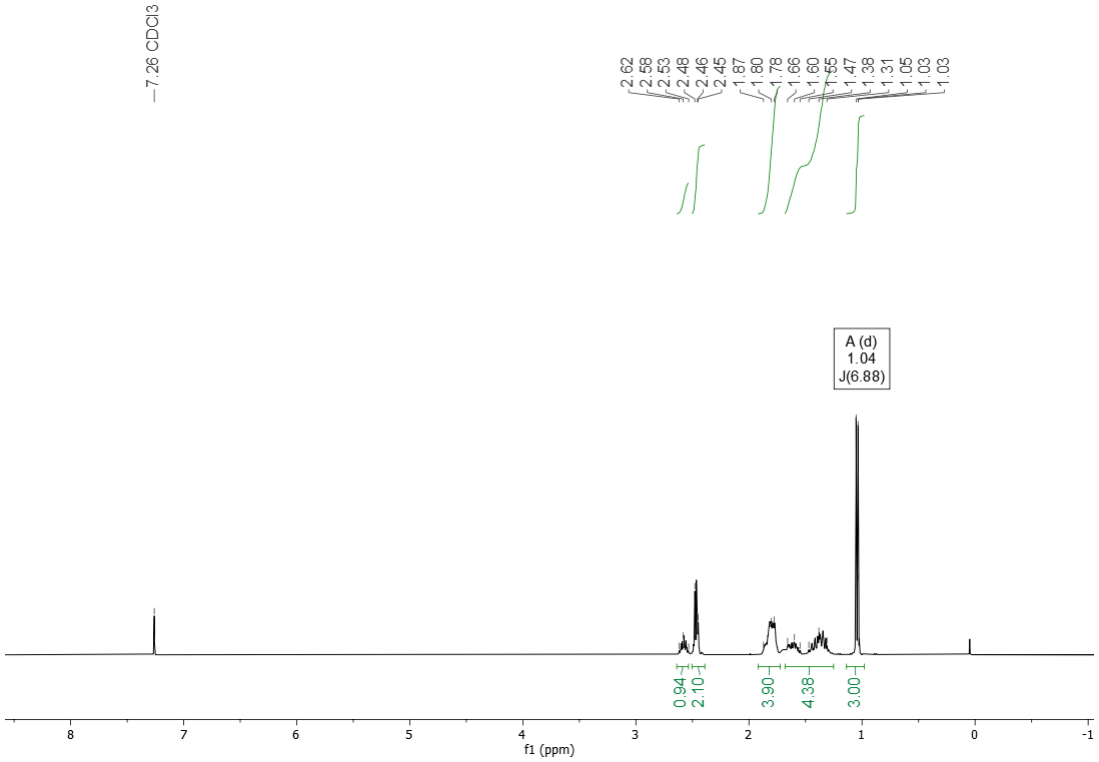


Figure 7.70: ^{1}H NMR (400 MHZ, $CDCl_3$) spectrum of 2-methylcycloheptanone.

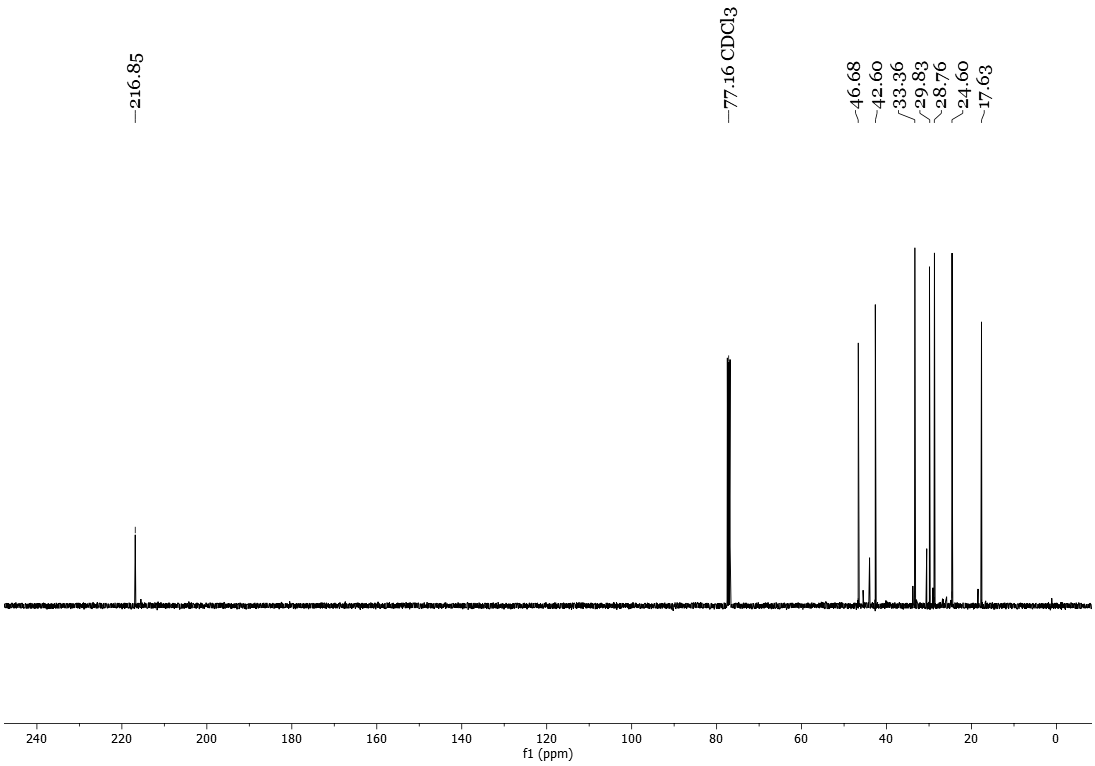


Figure 7.71: $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum of 2-methylcycloheptanone.

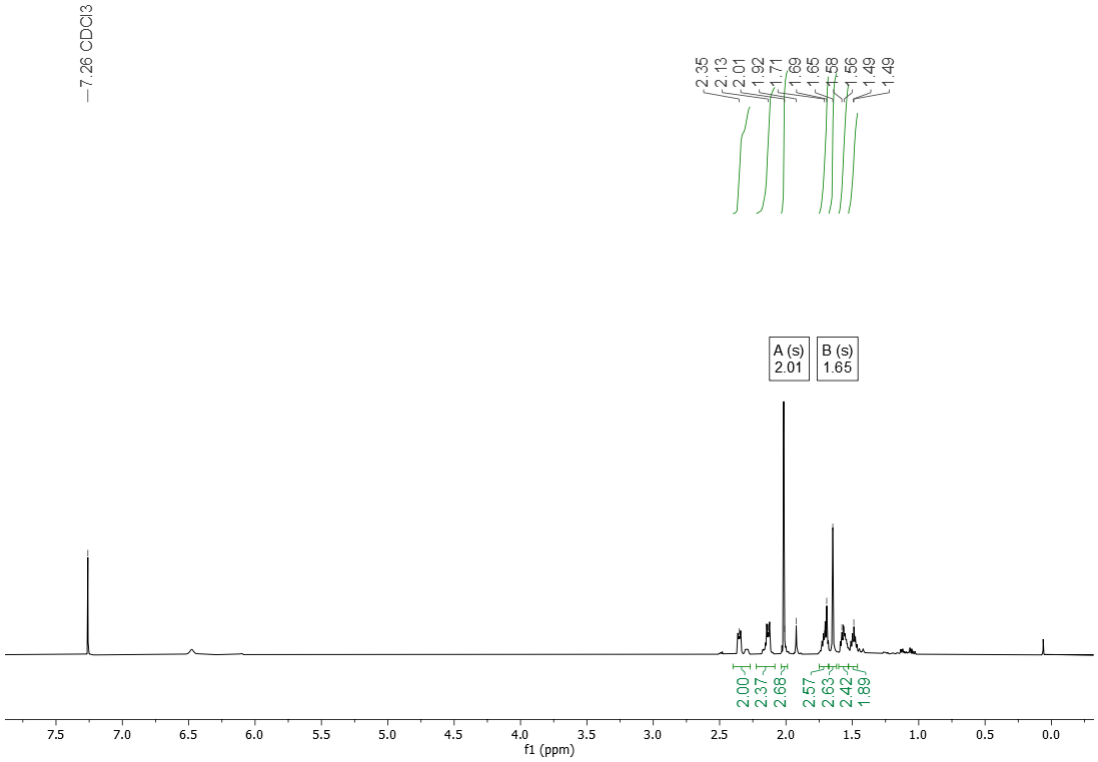


Figure 7.72: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of N-(2-methylcyclohept-1-en-1-yl)acetamide (**S4.1**).

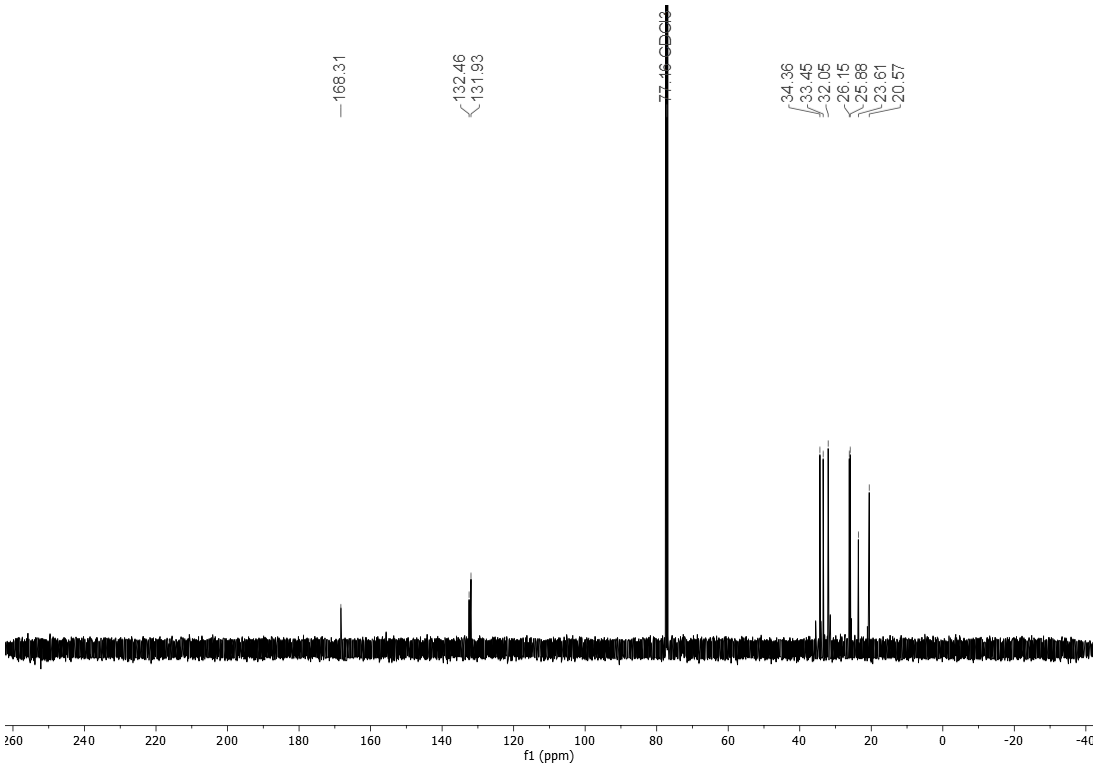


Figure 7.73: $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum of N-(2-methylcyclohept-1-en-1-yl)acetamide (**S4.1**).

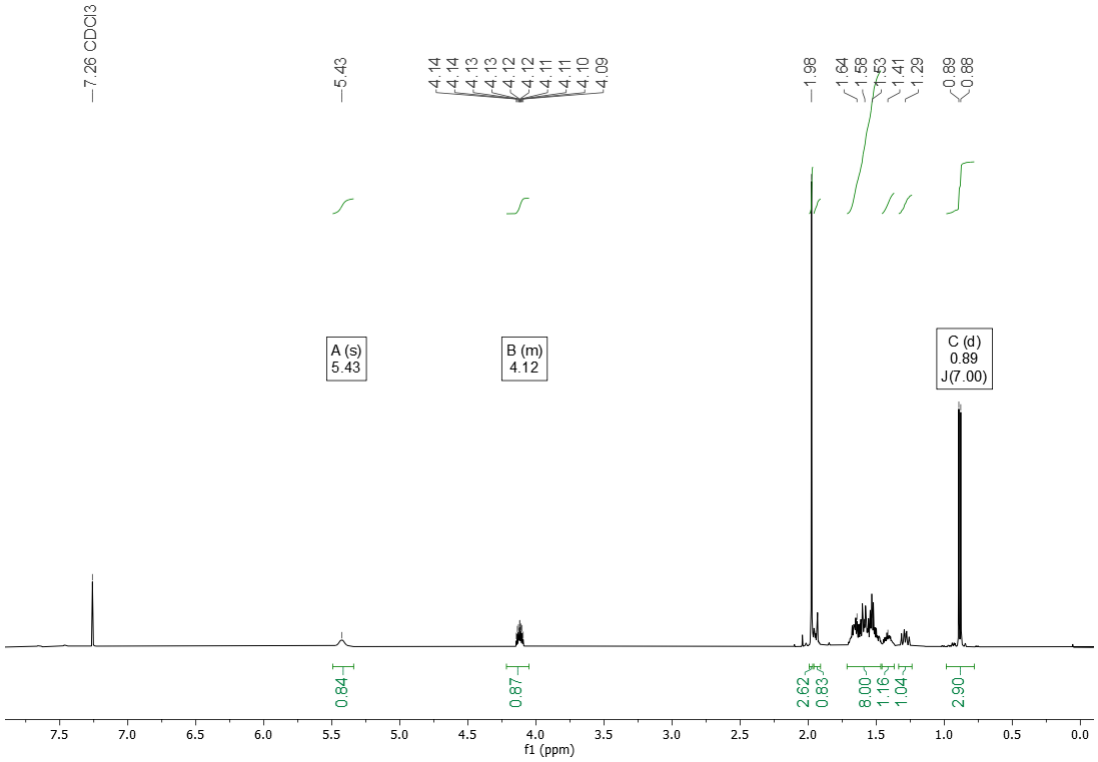


Figure 7.74: ^{1}H NMR (500 MHZ, $CDCl_3$) spectrum of N-(-2-methylcycloheptyl)acetamide (**P4.1**).

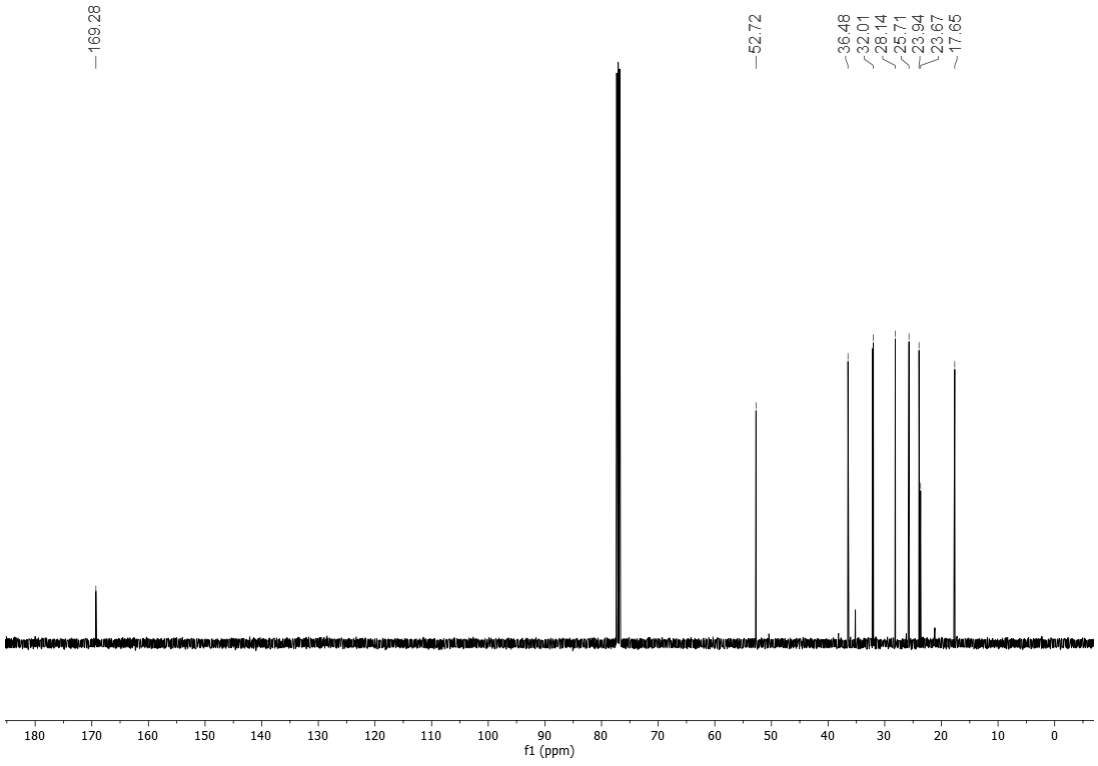


Figure 7.75: $^{13}C\{^{1}H\}$ NMR (126 MHZ, $CDCl_3$) spectrum N-(-2-methylcycloheptyl)acetamide (**P4.1**).