



Computational predictions and reactivity analyses of organic reactions

Ching Ching Lam
Trinity College

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

Supervisor: Prof. Jonathan M. Goodman
Yusuf Hamied Department of Chemistry
University of Cambridge

May 2024

Declaration

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.

Some of the work in this dissertation has been published in the following peer-reviewed journal articles:

- | | |
|-----------|---|
| Chapter 2 | M. M. Simpson, C. C. Lam , J. M. Goodman, S. Balasubramanian, Selective Functionalisation of 5-Methylcytosine via Organic Photo-redox Catalysis, <i>Angew. Chem. Int. Ed.</i> , 2023, 62, e202304756. |
| Chapter 3 | C. C. Lam , J. M. Goodman, Computational insights on the origin of enantioselectivity in reactions with diarylprolinol silyl ether catalysts via a radical pathway, <i>Org. Chem. Front.</i> , 2022, 9, 3730 |
| Chapter 4 | C. C. Lam , J. M. Goodman, CONFPASS: fast DFT re-optimisations of structures from conformation searches, <i>J. Chem. Inf. Model.</i> , 2023, 63, 4364–4375 |
| Chapter 5 | C. C. Lam , J. M. Goodman, Reaction dynamics as the missing puzzle: the origin of selectivity in oxazaborolidinium ion-catalysed reactions, <i>Chem. Sci.</i> , 2023, 14, 12355-12365 |
| Chapter 6 | C. C. Lam , J. M. Goodman, Every atom counts: Predicting sites of reaction based on chemistry within two bonds, manuscript submitted to <i>Digital Discovery</i> |
| N/A | D. Choudhury, C. C. Lam , N. L. Farag, J. Slaughter, A. D. Bond, J. M. Goodman, D. S. Wright, Suppressing cis/trans 'Ring-flipping' in Organoaluminium(III)-2-pyridyl Dimers – Design Strategies Towards Lewis Acid Catalysts for Alkene Oligomerisation, <i>Chem. Eur. J.</i> 2024, e202303872. |

Abstract

This thesis focuses on computational reactivity analyses and predictions for organic systems. The research began with studies on specific reactions using quantum mechanics and molecular mechanics simulations. Beyond looking at individual reactions, projects in method development were also initiated to set up a more efficient and automated procedure for tackling the challenges of exploring the conformational space, investigating the effect of reaction dynamics and suggesting possible reaction pathways.

Chapter 2 focuses on the selective pyridination of 5-methylcytosine. A new reaction mechanism that aligns with the experimental result was proposed. The variations in percentage yield upon changes in the substituent in the pyridine substrate can be explained by considering the thermodynamics and kinetics of the proton transfer step from cationic 5mdC^{•+} to neutral pyridine substrates. Substrates with an electron-withdrawing group are disfavoured thermodynamically, while substrates with an electron-donating group face a high kinetic barrier in a key proton transfer step.

Chapter 3 gives insights into the origin of enantioselectivity in 1,4-dicarbonyl synthesis reactions with diarylprolinol silyl ether catalysts via a radical pathway. A robust procedure has been developed for computational investigations of large and flexible chemical systems based on the conformation labelling system, ONIOM calculations and Python scripting. The change in enantiomeric excess due to variations in the catalyst can be explained based on conformational changes and structural deformations. In the enantioselectivity-determining radical addition step, the iminium in the lowest energy SR transition state (TS) takes up the conformation of the lowest energy ground state iminium (EE). The conjugated iminium in the SS TS adopts an EZ conformation to avoid potential structural deformations due to radical attacks from the more sterically hindered position. For systems with simpler catalysts, the iminium is EE for both, as the steric hindrance imposed by the substituent is not sufficient to cause this large structural deformation, and so the reaction shows poor enantioselectivity.

Chapter 4 is on the development of the CONFPASS (Conformer Prioritisations and Analysis for DFT re-optimisations) Python package. CONFPASS extracts dihedral angle descriptors from conformational searching outputs, performs clustering, and returns a priority list for density functional theory (DFT) re-optimisations. Evaluations were conducted with DFT data of the conformers for 150 structurally diverse molecules, most of which are flexible. CONFPASS gives a confidence estimate that the global minimum structure has been found, and based on our dataset, we can have 90% confidence after optimising half of the FF structures. Re-optimising conformers in order of the FF energy often generates duplicate results; using CONFPASS, the duplication rate is reduced by a factor of two for the first 30% of the re-optimisations, which includes the global minimum structure about 80% of the time.

Chapter 5 details the study of a group of oxazaborolidinium ion-catalysed reactions between aldehyde and diazo compounds. Their selectivity cannot be explained using transition state theory. VRAI-selectivity, developed to predict the outcome of dynamically controlled reactions, can account for both the chemo- and the stereo-selectivity in these reactions, which are controlled by reaction dynamics. Subtle modifications to the substrate or catalyst substituents alter the potential energy surface, leading to changes in predominant reaction pathways and altering the barriers to the major product when reaction dynamics are considered. In addition, this study suggests an explanation for the mysterious inversion of enantioselectivity resulting from the inclusion of an ortho ⁱPrO group in the catalyst.

Chapter 6 presents a method for predicting reaction sites based only on a simple, two-bond model. Machine learning classification models were trained and evaluated using atom-level labels and descriptors, including bond strength and connectivity. Despite limitations in covering only local chemical environments, the models achieved over 80% accuracy even with challenging datasets that cover a diverse chemical space. Whilst this simplistic model is necessarily incomplete, it describes a large amount of interesting chemistry and provides guidance on potential reaction pathways.

Acknowledgements

Firstly, I would like to express my sincere gratitude to Prof. Jonathan Goodman for accepting me into his research group. I have learnt tremendously through his continuous guidance and insightful support throughout my PhD. His limitless positivity and encouragements have really empowered me through some of the most difficult time of my PhD, from facing rejections and harsh reviewer comments to debugging the most random Python error which took more than a week to resolve. His attitude and ways of tackling challenges will be one of my biggest takeaways and will continue to inspire me in my journey beyond this PhD.

I wish to thank the members of the Goodman Group (in order of seniority and alphabetical order by last name: Kristaps Ermanis, Tim Allen, Guillermo Caballero, Alexander Howarth, Sanha Lee, Charles Chen, Marcus Wang, Mengman Wei, Katie Beckwith, Ruslan Kotlyarov, Henrique Marcon, Matthew Edge, Benji Rowlands) for the helpful discussion during the group meeting, technical support with software/programming, proofreading of this thesis and organisation of fun events during my time in the group.

For pyridination project (*ie* Chapter 2 of this thesis), I would like to thank Prof. Shankar Balasubramanian and Dr. Mathew Simpson for the sharing of data and experimental insights. I would also like to acknowledge Prof. Dominic S. Wright and Dipanjana Choudhury for giving me the opportunity to collaborate on the project in AI complex reactivity. I also hope to thank the IT support team at the department for helping me to clear up issues with the HPC.

I would like to acknowledge the funding supports (*ie* the Krishnan-Ang Studentship for Overseas Students in the Natural Sciences and the final term funding) from Mr Eashwar Viswanathan Krishnan, Ms Tzo Tze Ang and Trinity College. I feel very grateful to be part of Trinity. Thank you for supporting my student life and helping me settle into Cambridge during a global pandemic.

Finally, I want to thank the supports of my family and friends during this PhD journey. I would not be able to achieve this without their love and care.

Table of Contents

DECLARATION.....	2
ABSTRACT.....	3
ACKNOWLEDGEMENTS.....	5
GLOSSARY.....	9
INTRODUCTION.....	11
1.1 CHEMICAL REACTIVITY	11
1.2 THE ESTABLISHED APPROACH	12
1.3 THE BOTTLENECKS	14
1.3.1 <i>Exploring the conformational space</i>	14
1.3.2 <i>Beyond the transition state theory</i>	15
1.3.3 <i>Mechanism elucidation</i>	17
1.4 ROADMAP	20
PYRIDINATION OF 5-METHYLCYTOSINE: A COMPUTATIONAL STUDY ON REACTIVITY	22
2.1 BACKGROUND	22
2.2 COMPUTATIONAL METHODS	24
2.3 MECHANISTIC REVISION	25
2.3.1 <i>Step 2: Proton transfer</i>	27
2.3.2 <i>Step 3: C-C bond formation</i>	32
2.3.3 <i>Step 4: Hydrogen atom addition</i>	34
2.4 VARIATION IN PERCENTAGE YIELD.....	36
2.4.1 <i>Thermodynamic study</i>	36
2.4.2 <i>Kinetic study</i>	39
2.5 SUMMARY	40
REACTIONS WITH DIARYLPROLINOL SILYL ETHER CATALYST: INVESTIGATING LARGE AND STRUCTURALLY FLEXIBLE SYSTEMS	41
3.1 BACKGROUND	41
3.2 COMPUTATIONAL METHODS	43
3.3 MECHANISM AND THE ENANTIOSELECTIVITY-DETERMINING STEP	44
3.4 AUTOMATIONS AND CONFORMATION ANALYSES.....	47
3.5 ORIGIN OF THE ENANTIOSELECTIVITY	53
3.6 SUMMARY	61
CONFPASS: FAST DFT RE-OPTIMISATIONS OF STRUCTURES FROM CONFORMATION SEARCHES.....	62
4.1 BACKGROUND	62
4.2 DATA GATHERING	64

4.3 GENERATING CONFPASS PRIORITY LISTS.....	66
4.3.1 Dihedral angles descriptors	66
4.3.2 Clustering algorithms	72
4.3.3 Priority list generation methods.....	73
4.4 EVALUATING CONFPASS PRIORITY LISTS.....	75
4.4.1 Evaluation methods	75
4.4.2 Results.....	79
4.5 PREDICTING THE COMPLETION OF THE RE-OPTIMISATION PROCESS	84
4.6 CONFPASS AT WORK	89
4.7 COMPARISON WITH CREST	92
4.8 SUMMARY	96
OXAZABOROLIDINIUM ION-CATALYSED REACTIONS: REACTION DYNAMICS AS THE MISSING PUZZLE PIECE.....	97
5.1 BACKGROUND	97
5.2 COMPUTATIONAL METHODS	99
5.3 THE MECHANISM.....	102
5.4 PREDICTING SELECTIVITY WITH VRAI-SELECTIVITY.....	104
5.5 IMPLICATIONS FOR SELECTIVITY	113
5.5.1 Stereoselectivity.....	113
5.5.2 Chemoselectivity.....	117
5.6 SUMMARY	119
PREDICTING SITES OF REACTION BASED ON CHEMISTRY WITHIN TWO BONDS	120
6.1 BACKGROUND	120
6.2 DATA.....	122
6.2.1 Data gathering	122
6.2.2 Data processing.....	123
6.3 REACTIVE SITE PREDICTION MODEL.....	125
6.3.1 Descriptor.....	125
6.3.2 Label	127
6.3.3 Algorithm.....	128
6.4 RESULTS AND DISCUSSION	129
6.4.1 Type A: the first-year dataset	132
6.4.2 Type B: the [3+2] cycloaddition and the Diels-Alder dataset	135
6.4.3 Type C: the RGD1 and the green dataset.....	137
6.4.4 Model generalisability	139
6.5 SUMMARY	140
RECAPITULATION, CONCLUSIONS AND FUTURE WORKS	141
7.1. CONCLUSIONS.....	141

7.2. FUTURE WORKS	143
APPENDIX	144
NOTES ON THEORIES	144
NOTES ON CALCULATIONS	148
APPENDIX TO CHAPTER 2.....	149
2.A Data availability	149
2.B Overall energy diagram	150
APPENDIX TO CHAPTER 3.....	151
3.A Data availability	151
3.B Dihedral angle distribution studies.....	152
3.C Benchmarking	154
3.D Distortion-interaction analyses	157
3.E Non-covalent interaction analyses.....	158
3.F Overall energy diagram	159
APPENDIX TO CHAPTER 4.....	160
4.A Data and scripts.....	160
4.B Generating CONFPASS priority lists	162
4.C Post-processing procedures.....	164
4.D Evaluating CONFPASS priority lists.....	167
4.E Benchmarking: different levels of theory	168
4.F Predict the completion of the re-optimisation process	173
4.G Comparison with CREST	177
APPENDIX TO CHAPTER 5.....	186
5.A Data availability	186
5.B Calculation breakdowns	188
5.C Benchmarking	200
5.D Understand selectivity	203
APPENDIX TO CHAPTER 6.....	206
6.A Data availability	206
6.B Bond strength descriptor	212
6.C Model evaluation	213
REFERENCE	216

Glossary

%Conf	Percentage confidence in CONFPASS: equivalent to the percentage of true prediction
5mdC	5-methyl-2'-deoxycytidine
B3LYP	Becke exchange with Lee-Yang-Parr correlation
B3LYP-D3	B3LYP with D3 dispersion correction
COBI	Chiral oxazaborolidinium ion
CONFPASS	Conformer Prioritisations & Analysis for DFT re-optimisations; a Python package for fast exploration of conformational space
CREST	Conformer-Rotamer Ensemble Sampling Tool
DFT	Density Functional Theory
DHP	1,4-dihydropyridine
ee	Enantiomeric excess
FF	Force field
GaussianNB	Gaussian Naive Bayes
InChI	The International Chemical Identifier
INT	Intermediate
IRC	Intrinsic reaction coordinate
KDE	kernel density estimation
KNN	K-nearest neighbour classification
LR	logistic regression
<i>m</i>	Total number of conformers from conformational searches at FF level
M06	Minnesota 06 functional
MD	Molecular dynamics
MM	Molecular mechanics
<i>n_clusters</i>	Number of clusters
ONIOM	A hybrid QM/MM method, short for ' O ur own N -layered Integrated molecular O rbital and M olecular mechanics'
OTDS	Oxy-thexyldimethylsilyl group
OTMS	Oxy-trimethylsilyl group
PES	Potential energy surface
P_{GMT}	The proportion of conformers need to be optimised in order to obtain the lowest energy conformer at the DFT level; a CONFPASS evaluation metric
P_{overall}	$P_{\text{overall}} = P_{\text{GMT}} + 5(\Delta a_{\text{bins}})$; a CONFPASS evaluation metric
Q	Key parameter in the CONFPASS <i>pipeline-mix</i> method: the ratio of the <i>pipeline-x</i> content in the new <i>pipeline-mix</i> priority list
QM	Quantum mechanics
QRC	Quick reaction coordinate

r_{bins}	The ratio of populated bins for a particular selection of conformers from partial optimisations; the number of bins is the number of distinct structures at DFT level; a parameter to obtain CONFPASS evaluation metrics
RF	Random forest
RGD1	Reaction Graph Depth 1 from Savoie <i>et al.</i> (<i>Sci Data</i> , 2023 , 145)
RMSD	Root-mean-square deviation
r_{opt}	The ratio of the number of DFT re-optimised conformers over the total number of conformers at FF level; a CONFPASS parameter
r_{opt,conf=c}	The r _{opt} at which %Conf no longer deviates below c in the %Conf vs r _{opt} plot; a CONFPASS parameter
SMILES	Simplified molecular-input line-entry system
TS	Transition state
TST	Transition state theory
UFF	Universal force field
x	Key parameter in the CONFPASS <i>pipeline-x</i> method: $n_clusters=m*x$
Δa_{bins}	The area between the DFT list and proposed priority list curve in a r _{bins} vs r _{opt} plot; a parameter to obtain CONFPASS evaluation metrics
η	Range
θ	Dihedral angle value
μ	Mean
σ	Standard deviation
χ	Mole amount; $\chi=\exp(-\Delta G/RT)$; a CONFPASS parameter
χ_{new}	Mole amount of the last optimised conformer; a CONFPASS parameter
ωB97X-D	Head-Gordon's Long-range corrected hybrid density functionals with dispersion corrections

Chapter 1

Introduction

1.1 Chemical reactivity

The research on chemical reactivity is about interpreting and rationalising the reaction outcomes.^{1–3} Thorough and comprehensive analyses of the reaction outcomes provide a foundation for building predictive models and guiding future experiments, which are critical for deriving better strategies and driving the development cycle in synthetic chemistry (Figure 1.1.A).⁴

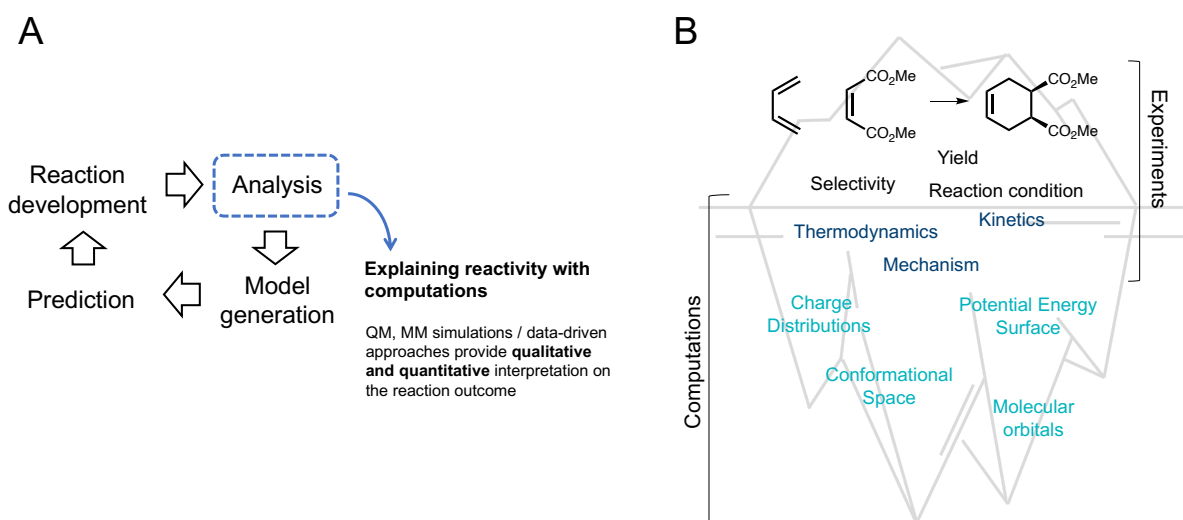


Figure 1.1. Investigating chemical reactivity. A) The development cycle in synthetic chemistry (QM = quantum mechanics, MM = molecular mechanics). The diagram is proposed by Duarte *et al.*⁴ B) A chemical reaction mirrors an iceberg.

A chemical reaction mirrors an iceberg (Figure 1.1.B). The reaction outcomes — product selectivity, yield, and the optimal range of reaction conditions — are relatively straightforward to obtain from experiments and represent the tip of the iceberg. However, to explain the reactivity, one must look beyond the surface level. This means exploring the potential energy surface in searching for the most plausible reaction mechanism and examining contributions from various steric and electronic factors. Going beneath the surface through experiments is notably challenging and time-consuming. This is where computations, from the well-established simulation approaches to data-driven methods, come into play to provide qualitative and quantitative interpretations of reaction outcomes.

1.2 The established approach

The most well-established computational method for explaining reactivity is through simulations with quantum mechanics (QM) and molecular mechanics (MM) approaches.^{3,5–8}

Density functional theory (DFT) is one of the most frequently used QM methods for studying organic reactions. Built on the idea that the ground state of a chemical system is determined by its electron density, DFT calculations approximate the solution to Schrodinger's equation.⁷ The most frequently used DFT functionals within the organic chemistry research community are B3LYP-D3, ω B97X-D and M06-2X for their applicability to common organic systems.⁹ The work into deriving better DFT functionals (*ie* universal and accurate in terms of energy) remains an important field of research. The list of available functionals has continued to grow in the last ten years. In 2017, Head-Gordon *et al.* benchmarked over 200 functionals on a database of 5000 molecules regarding energy prediction.¹⁰ The wide range of functionals can support the demand of applications on different chemical systems.¹¹

While DFT provides accurate structural and energetics insights, MM approaches allow fast exploration of the chemical space of the organic system.⁶ In MM approaches, molecules are treated as an assembly of particles connected by springs obeying simple harmonic motion principles.^{7,12} A molecular force field (FF) describes these forces between the particles or atoms. The energy function is derived by summing the contributions from various covalent and non-covalent components, including bond stretching, bending, dihedral angle rotation, van der Waals and electrostatic interactions. FFs are derived empirically and parameters are chosen to allow best alignment with the experiment. The applicability domain of an FF may vary depending on the parametrisations and the range of molecular systems for which it has been validated. FFs, such as MM2,¹³ MM3,¹⁴ MMFF (Merck molecular force field)¹⁵ and OPLS4^{16,17}, are designed for studying organic molecules and are widely used within the research community.¹⁸ Optimisations and energy calculations at the FF level are computationally cheap. Therefore, conformational searches tend to be conducted with FF for the speed of the calculations.^{19,20}

Computational investigations on chemical reactivity with simulations involve several steps (Figure 1.2).^{1,4} The first step usually involves exploring the conformational space of the chemical systems at the FF level. Conformational searching is especially important for a structurally complex and flexible molecule.^{21–23} Sequentially, the FF structures are re-optimised using DFT, which is more accurate but comparatively slow. Further calculations and analyses, such as transition state (TS) optimisations and non-covalent interactions analyses,²⁴ can then be conducted to rationalise the reactivity and selectivity.

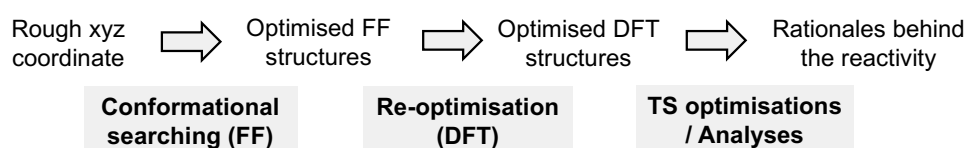


Figure 1.2. The computational workflow to understand chemical reactivities with MM and QM simulations.

One influential study on chemical reactivity based on simulations was conducted by Houk and List in 2003.^{22,25} They established a model for explaining the catalytic activities of proline in stereoselective aldol reactions (*ie* the Houk-List model) with DFT calculations. Stereochemical outcomes (Figure 1.3) of the reaction depend on the face of the aldehyde attack (*re/si*) and the relative orientation of the enamine and carboxylic acid group (*syn/anti*). The model suggests that the most kinetically favourable TS should have minimal steric clashes, strong electrostatic stabilisation from non-covalent interactions and a planar enamine structure (*ie* greater nucleophilicity at the enamine C=C bond).²⁶

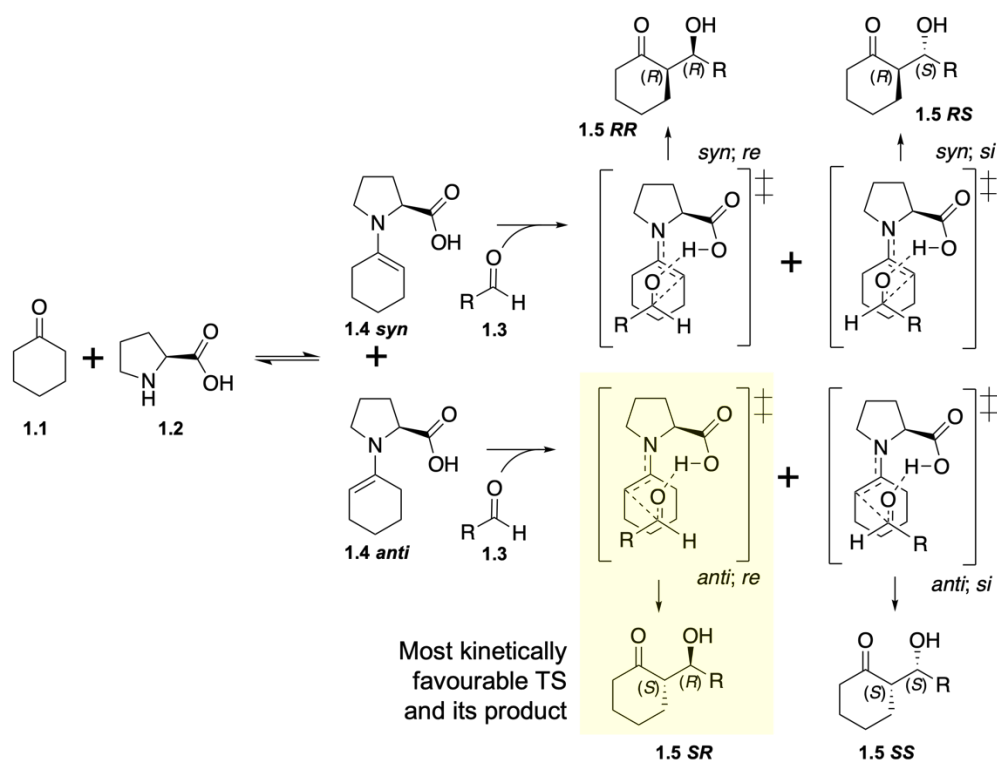


Figure 1.3. The Houk-List model: Possible stereochemical outcomes in asymmetric aldol reactions are presented with the associated TSs.⁹

The model demonstrates decent transferability on the stereochemical prediction for other proline-catalysed processes, including Michael and Mannich reactions.²⁷ The most influential aspect of the study was not the final model but rather the systematic approach developed for predicting and explaining reactivity based on quantitative analyses.⁴ Based on MM and QM simulations, models for rationalising and predicting selectivity and catalytic activity have been developed for various chemical systems, such as chiral phosphoric acid (CPA)-catalysed nucleophilic addition on iminium,²⁸ the bifunctional iminophosphorane (BIMP)-catalysed synthesis of cyclohexenones²⁹ and aza-Diels–Alder reactions on complex substrates.^{30,31}

1.3 The bottlenecks

The established approach with QM and MM simulations is not perfect. In the following section, several bottlenecks inherent in the computational workflow in Figure 1.2 are discussed.

1.3.1 Exploring the conformational space

Molecular structure and conformational space are important to organic chemists in understanding the structure and reactivity of a molecule.^{1,32} How many conformations do we need to consider to get a clear idea of the properties of a molecular system? This is central to the analysis of structure and reactivity. Recently, the potential of conformational flexibility as a tool in controlling the yield of the desired product and designing better catalysts has also been explored.³³ Several reports have highlighted the benefits of conformational flexibility in catalyst designs, which include enhancing non-covalent interactions and improving substrate applicability.^{34–36} However, the reactivity of a conformationally flexible molecule is difficult to investigate due to the vast conformational space and the mechanistic complexity.^{23,37,38} Previously, Lledos *et al.* have pointed out that the error range associated with the conformational diversity of the Pd catalyst can be up to 11.4 kcal mol⁻¹ in computational studies of Suzuki coupling reactions.²³

In the established computational workflow for investigating reactivities with simulations (Figure 1.2), DFT optimisations of the FF level structures are a bottleneck. Conformer generators, such as MacroModel,³⁹ OMEGA^{40,41} and RDKit⁴², can analyse the results of conformational searches and remove duplicate structures and high-energy conformers.^{43–45} The search for even better approaches is ongoing. CREST (Conformer-Rotamer Ensemble Sampling Tool) is a state-of-art program which performs conformational searching calculations with the iterative meta-dynamics with genetic crossing (iMTD-GC), based on a semi-empirical method (GFN2-xTB).^{46–48} At the end of the process, conformers are identified from rotamer ensembles based on RMSD, energy and rotational constant thresholds. Even using such procedures, the conformational searching output files of structurally flexible molecules may include hundreds of relevant structures.⁴⁹ Optimising all the conformers from the output files at the DFT level requires a huge amount of computational time. Sometimes, distinct FF structures may end up having the same conformation at the DFT level (Figure 1.4). However, re-optimising just a few conformers with low FF energy is also insufficient. Thorough explorations of the conformational space are critical for understanding reactivities.¹⁹ The correlations between the FF or semi-empirical potential energy surface (PES) and the DFT PES are difficult to predict for individual molecules, which has been reported in benchmarking studies.^{50,51} As is well known, a relatively high energy conformer (e.g. $\Delta U(\text{FF}) = 4 \text{ kcal mol}^{-1}$ compared to the lowest energy structure) at the FF level might be the global minimum at the DFT level.⁵²

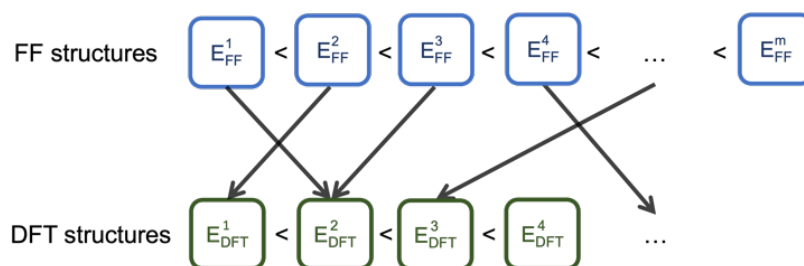


Figure 1.4. Sometimes, force field structures ordered by energy are re-optimised at a DFT level to a different order and a different number of structures.

There is a need for a better approach to explore the conformational space of organic systems more efficiently and effectively with minimum computational cost.

1.3.2 Beyond the transition state theory

Computing organic stereoselectivity is now a standard part of synthetic and mechanistic chemistry.¹ In this process, the key intermediates and TSs structures are optimised for calculating thermodynamic and kinetic parameters (Figure 1.5).^{1,53} In general, either the pathway with the lowest energy transition state leads to the preferred product, or else the thermodynamically favourable product is formed.⁵⁴ This approach proves to be effective in explaining the reactivity of many organic reactions,⁵ including the examples discussed in section 1.2 of this chapter.

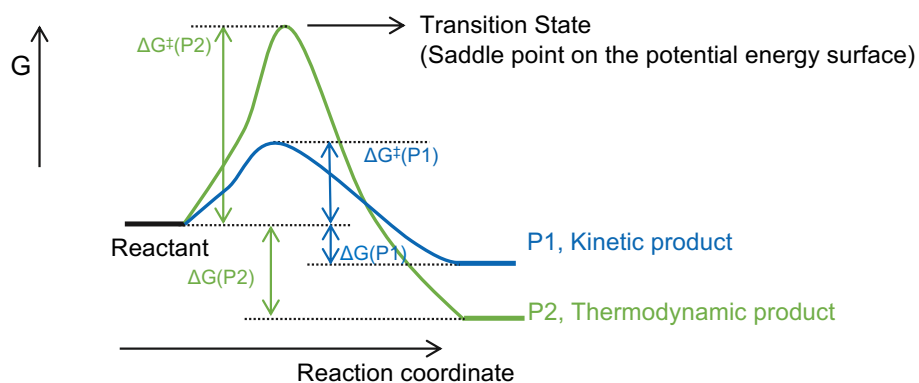


Figure 1.5. Predicting selectivity based on kinetic and thermodynamic parameters.

There are also reactions where the product selectivity is strongly influenced by reaction dynamics. In these cases, calculating TS energies by itself is insufficient to understand the outcome.^{55–61} Such reactions tend to have bifurcating reaction surfaces⁵⁵ or involve processes with a shallow intermediate that connects two TS, where the first TS is much higher in energy than the second one (Figure 1).^{60,61} The role of reaction dynamics has been largely overlooked. One possible reason might be the high computational cost. Traditionally, expensive molecular dynamics (MD) simulations need to be run to produce a large number of reaction trajectories to confirm the effect of reaction dynamics.⁶²

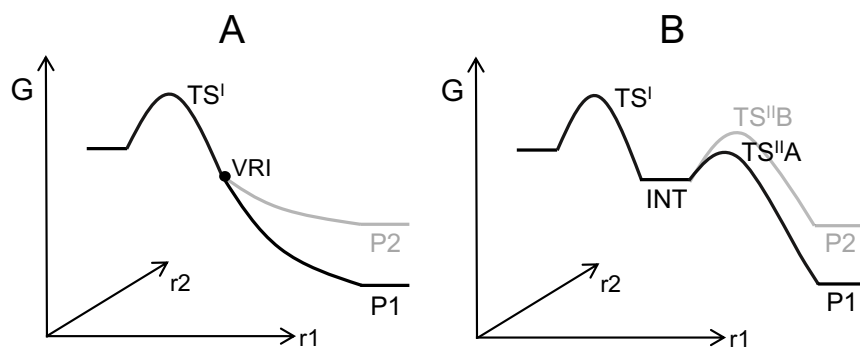


Figure 1.6. Examples of processes with dynamically controlled selectivity: A) a bifurcating reaction with a valley-ridge inflection point (VRI) on the potential energy surface⁵⁵ and B) a reaction with a shadow intermediate where the sequential TSs are low in energy compared to the first TS.^{60,61}

Recently, the Goodman group developed an algorithm, VRAI-selectivity, to analyse and predict selectivity outcomes of dynamically controlled reactions straightforwardly without expensive computations.^{63,64} VRAI-selectivity was developed and tested on about 60 reactions from 25 papers published from 2003 to 2019. Examples include the work on bi- and tri-pericyclic reactions from Houk *et al.*,^{65,66} the study of bifurcating Rh-carbenoid C–H insertions reactions from Tantillo *et al.*⁶⁷ and the publication on dynamic effects in alkene hydroboration from Singleton *et al.*⁶⁰ The size of these chemical systems was surveyed and presented in Figure 1.7 as a histogram. The number of atoms involved in the reactions ranges from 10 to 50, with a median of 25. Recent research has also revealed a growing number of complex organic reactions, in which a significant amount of diligent effort and computational resources have been dedicated to verify the effect of reaction dynamics. Houk *et al.* have just reported the discovery of tetra-pericyclic cycloadditions.⁶⁸ Ess *et al.* have detailed the dynamic effects of organometallic reactions in their published works.⁶⁹ The research in terpene forming reactions, which are known to be dynamically controlled, is ongoing in the community.^{70–72} Will VRAI-selectivity be able to confirm the reaction dynamic effects in more complex chemical systems? This is worth investigating.

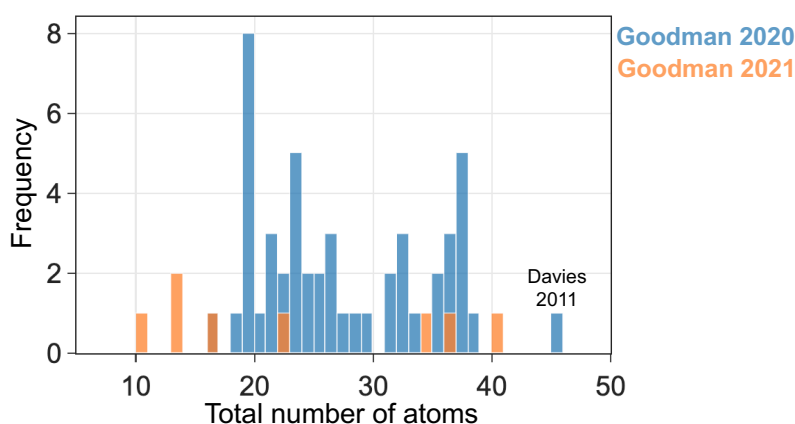


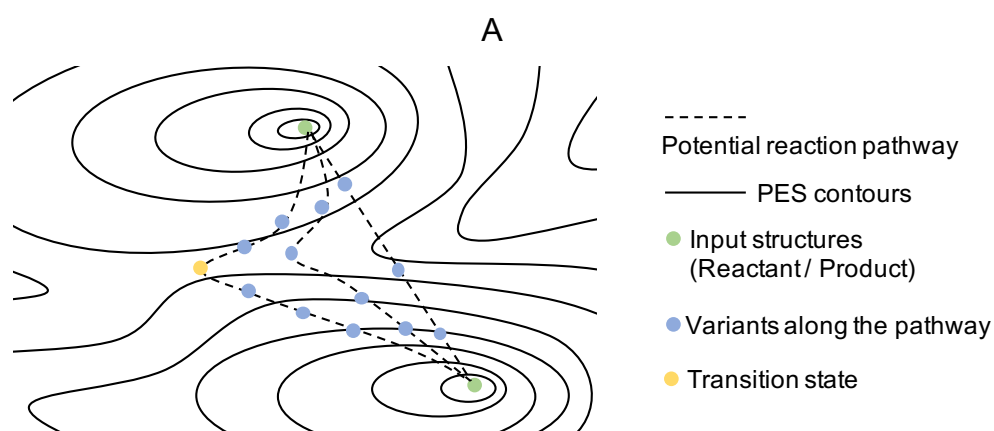
Figure 1.7. Distribution of the size of the chemical systems covered in the development of VRAI-selectivity (Goodman 2020 and 2021).^{63,64} The reaction with the greatest number of atoms comes from the work of Davies *et al.*⁷³

1.3.3 Mechanism elucidation

Elucidating the reaction mechanism is a long-standing challenge. Most conveniently, chemists often relied on intuitions and knowledge of the chemistry of the reaction.⁷⁴ Unobvious but critical pathways and TS orientations might be missed out with this approach. However, thorough exploration of the potential energy surface by calculations can also be very time-consuming. How can we achieve a balance between leveraging existing knowledge and conducting a comprehensive exploration of the PES in mechanism elucidations? Various innovative approaches have been developed to automate the processes of mechanism elucidation and identify the intermediates or TSs with reduced efforts.⁵³

When the reaction mechanism is completely unknown, thorough and time-consuming examinations of all the possible reaction pathways may be essential.⁵³ Methods based on MD simulations were developed to tackle such challenges by utilising the random nature of reaction collisions and allowing a thorough exploration of PES. Examples include the transition state search using chemical dynamics simulations (TSSCDS) from Martinez-Nunez⁷⁵ and the *ab initio* nanoreactor method⁷⁶ from Martinez *et al.* In these approaches, key intermediates and TS structures are estimated by analysing the time evolution of chemical systems at the atomic level along the reaction trajectory.

More commonly, we will have some level of understanding about the reactions in terms of the potential intermediates and products. Many methodologies are designed with this awareness in mind to automate the traditional DFT bond scanning process. The growing string method (GSM) is one of the examples (Figure 1.8.A).^{77–79} Given initial structure geometries (*ie* reactants and products), the GSM explores the PES by automatically generating structures with progressively altered geometries. The process is repeated along different paths until a TS structure is found.⁷⁹ Alternatively, the artificial force induced reaction (AFIR) method⁸⁰ involves distorting the PES by adding an artificial force. The distortion leads to a downhill pathway to the products of the reaction. The TS position can be approximated based on the path by examining the gradient of the distorted PES.



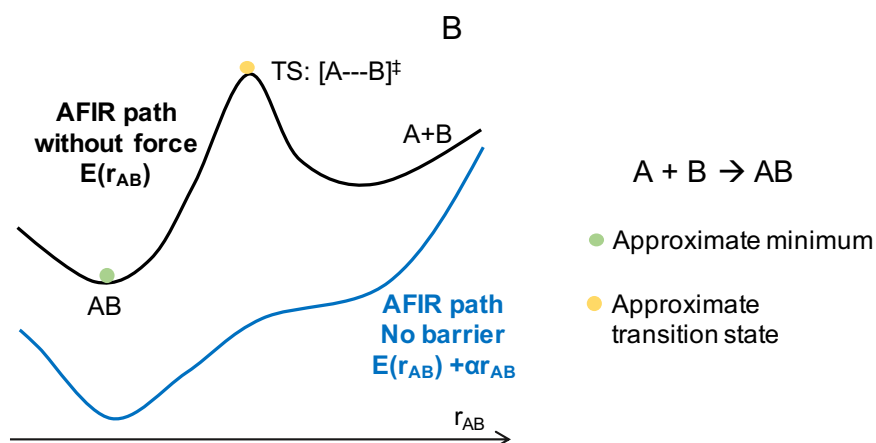


Figure 1.8. Exploring the PES: A: The double-ended growing string method (GSM);⁷⁹ B: The artificial force induced reaction (AFIR) method.⁸⁰ $E(r_{AB})$ is the PES along path r_{AB} and α is the artificial force constant.

Decades of experimental and computational research have led to the generation of a huge amount of data. Utilising these existing data allows us to stand on the shoulders of giants and derive reaction pathways of known reactions.^{53,81,82} A way of computerising chemical transformations is through molecular graphs, where atoms are nodes and bonds are edges. Rules and reaction templates are implemented on the molecular graphs as matrix transformations. This approach has been integrated into many automated systems for deriving reaction pathways, such as the Reaction Mechanism Generator from West *et al.*⁸³ and the Automated Calculation of Reaction Energy Profiles (autodE) programme from Duarte *et al.*⁸⁴ Recently, there is a growth in the application of machine learning models in predicting organic transformation.⁶² Multiple research groups have explored the potential of neural network models in ranking potential transformations or mechanism.^{85–88} Lately, chemists have ventured into adapting language models for predicting reactivity, utilising a text-based representation, such as SMILES and InChI,⁸⁹ as input.⁹⁰ The T5Chem⁹¹ model developed by Zhang *et al.* and the Molecular Transformer⁹² model from Lee *et al.* are some recent advancements in the field.

Data-driven approaches typically offer fast prediction time but demand a substantial volume of high-quality data for effective model generation and evaluation.^{93,94} Simulations, such as the MD, GMS and AFIR approaches, can handle scenarios with little existing chemical knowledge but often require high computational costs.⁵³ All approaches for elucidating mechanisms present a trade-off (Figure 1.9), which is unavoidable. A possible avenue for future research can be focusing on delving deeper into the methodologies to gain a better understanding of their limits and potential.^{95–97} This will aid in determining the most opportune moments for the methodological application.

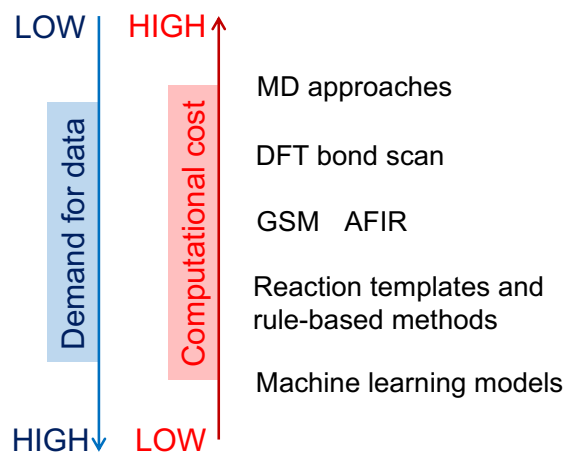


Figure 1.9. All approaches for elucidating mechanisms present a trade-off between demand for data and computational cost for making predictions. (DFT = density functional theory, MD = molecular dynamics, GSM = growing string method, AFIR = artificial force induced reaction)

1.4 Roadmap

This thesis revolves around computational predictions and reactivity analyses of organic reactions. The research begins with investigations on individual icebergs of chemical reactions. Beyond having an in-depth understanding of a select few reactions, method development projects are initiated to comprehend and explain the reactivity efficiently and effectively. (Figure 1.10)

Chapter 2 focuses on rationalising the reactivity trend in a series of 5-methylcytosine pyridination reactions using the established approach with MM and QM calculations (Figure 1.2).

Chapter 3 presents the computational insights on the origin of enantioselectivity in reactions with diarylprolinol silyl ether catalysts. This is a complex organic system with a high degree of conformational flexibility. The established computational workflow in Figure 1.2 alone was not sufficient to tackle the chemical system. Here, an automated computational procedure was introduced to analyse the outputs of conformational searching and obtain low-energy structures for this organocatalytic system.

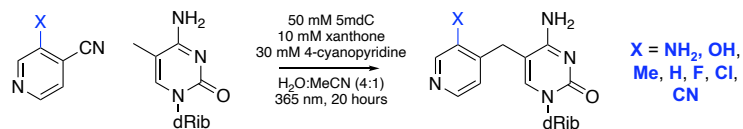
In the search for a universal method to rapidly explore conformational space and building from the work in Chapter 3, the CONFPASS project was initiated. The investigation is described in Chapter 4. CONFPASS is a Python package that employs clustering algorithms and machine learning classification models to streamline the analyses of conformational searching outputs and guide DFT re-optimisations.

Sometimes, thorough conformational explorations of the possible TS geometries are not enough to predict the selectivity of an organic reaction. Chapter 5 of this thesis illustrates such a scenario. The selectivity in a group of oxazaborolidinium ion-catalysed reactions between aldehyde and diazo compounds cannot be explained using transition state theory. VRAI-selectivity can account for both the chemo- and the stereo-selectivity in these reactions, which are controlled by reaction dynamics.

Chapter 6 presents a method for predicting reaction sites based only on a simple, two-bond model. This is a simplistic approach. The investigation is conducted to understand the limits of data-driven approaches while provide insights into the molecule reactivity and hinting at the possible transformations.

Chapter 2

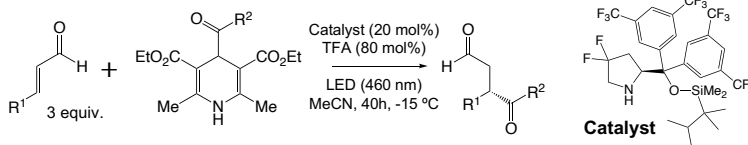
Pyridination of
5mdC
(Angew. Chem. Int. Ed.,
2023, e202304756)



Applying the established
computational workflow
to understand reactivities
in new reactions

Chapter 3

1,4-dicarbonyl
synthesis
(Org. Chem. Front.,
2022, 9)



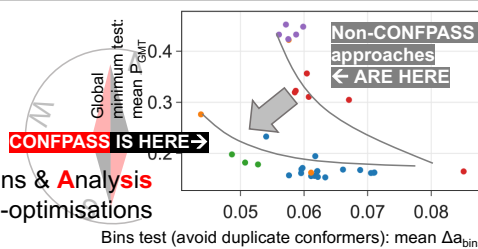
Exploring the
conformational
space efficiently
and effectively

Chapter 4

CONFPASS
(J. Chem. Inf.
Model., 2023, 63)

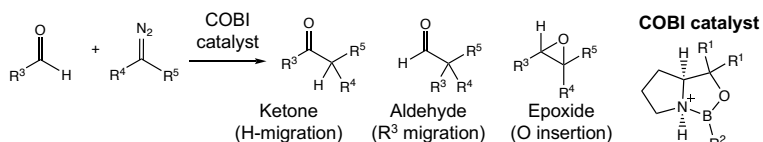
- ✓ Automated, rapid and robust
- ✓ Focus on organic molecules
- ✓ Guide your calculations with enhanced confidence

Conformer Prioritizations & Analysis
for DFT re-optimisations



Chapter 5

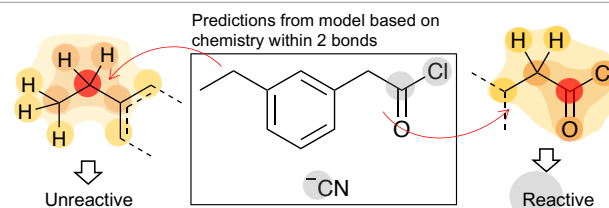
COBI-catalyzed
reactions
(Chem. Sci., 2023, 14)



Reaction dynamics
as the missing
puzzle piece in
explaining reactivity

Chapter 6

Predicting reactivity
based on chemistry
within 2 bonds



Towards effective
mechanism
elucidations

Figure 1.10. The roadmap of this PhD journey: the graphic abstract for chapter 2-6.

Chapter 2

Pyridination of 5-methylcytosine: A computational study on reactivity

2.1 Background

This chapter describes the computational investigation of a series of 5-methylcytosine pyridination reactions (Figure 2.1). This is a collaborative project with the Balasubramanian group at Cambridge, where all the computational works were carried out by the author of this thesis. All experimental results relevant to this chapter are presented in Figure 2.1. The experiments were conducted by Dr. Mathew M. Simpson under the supervision of Prof. Shankar Balasubramanian. Details of the experiments can be found in this publication.⁹⁸

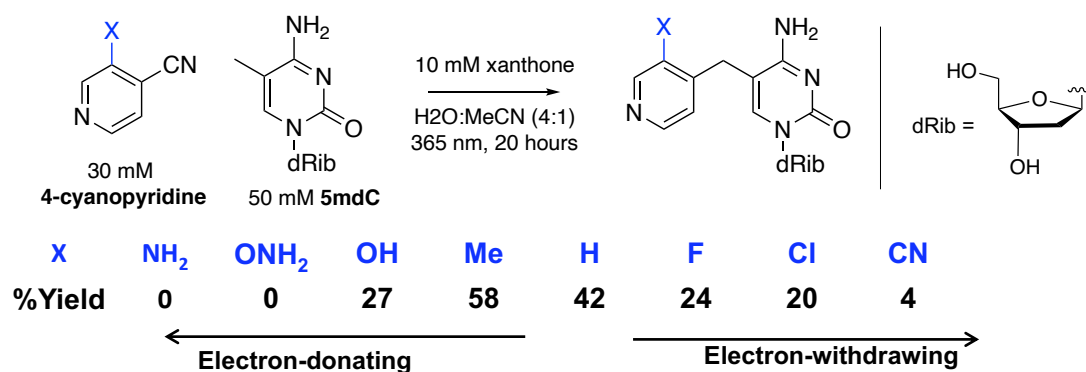


Figure 2.1. Pyridination of 5-methylcytosine experimental results. The percentage yield of the reactions was determined with mass spectrometry.

5mC (5-methyl-2'-deoxycytidine) is often referred as the '5th base' due to its abundance in human genome, where about 4% of cytosine are methylated.⁹⁹ The selective chemical modification of the 5mC methyl group has been a long-standing challenge. These reactions have the potential to improve the detection and accelerate biomedical research on the 5mC epigenetic feature. Furthermore, pyridine scaffolds are essential building blocks for various natural products, pharmaceuticals and organic materials.^{100–102} Selective pyridination of C-H bonds offers a highly strategic avenue for constructing complex motifs and provides a stepping stone for sequential reactions.¹⁰³

Experimental trials were conducted with various 4-cyanopyridine derivatives, where the substituent at position C3 (referred to as 'the substituent' or -X in the text below) was varied (Figure 2.1). Increasing the electron-donating or electron-withdrawing strength of the substituent led to a drop in percentage yield. The experimental results imply that the reactivity is likely to be governed by multiple factors besides the electronic properties of the ring systems.

With computational approaches, this project aims to

1. establish a mechanism for pyridination of 5mdC.
2. explain the variation in percentage yield between the pyridine substrates.

2.2 Computational methods

Conformational searching calculations were conducted with MacroModel (v11.7) with Maestro at the MMFF¹⁵ level for all monomolecular and multimeric systems. A setting of 1000 as the maximum number of steps were used. Conformers within an energy window of 5.0 kcal mol⁻¹ (21 kJ mol⁻¹) were saved and re-optimised with DFT calculations. The default setting, 0.5 Å, was used for the RMSD cut-off to filter conformers with similar structures.

For the multimeric systems, up to 100 conformers may be obtained from a conformational searching calculation. Re-optimising all conformers would be very computationally expensive. Hence, if a conformational searching output file contained more than 30 conformers, a Python script would be used to select structures for re-optimisations (github.com/Goodman-lab/CSearch_alignmol). The input to the script is the conformational searching output files in xyz and sdf format, which contain coordinates of conformers in order of the relative U at the MMFF level. The script selects every *n*-th structure and the structural outliers. The value of *n* was chosen to be between 2 to 5 in this project to keep the sample size of the selected conformers under 30. The structural outliers are identified based on root-mean-square deviation (RMSD) measurements using the RDkit.AlignMol function. A large RMSD value implies that the two structures are distinctively different. Pairwise structural similarity comparisons are conducted for each conformer by measuring the RMSD with other conformers from the conformational searches. This leads to an array of RMSD values for each conformer. The average RMSD is calculated for each conformer based on the values in the RMSD array. A conformer is identified as a structural outlier when it meets one of the two criteria below:

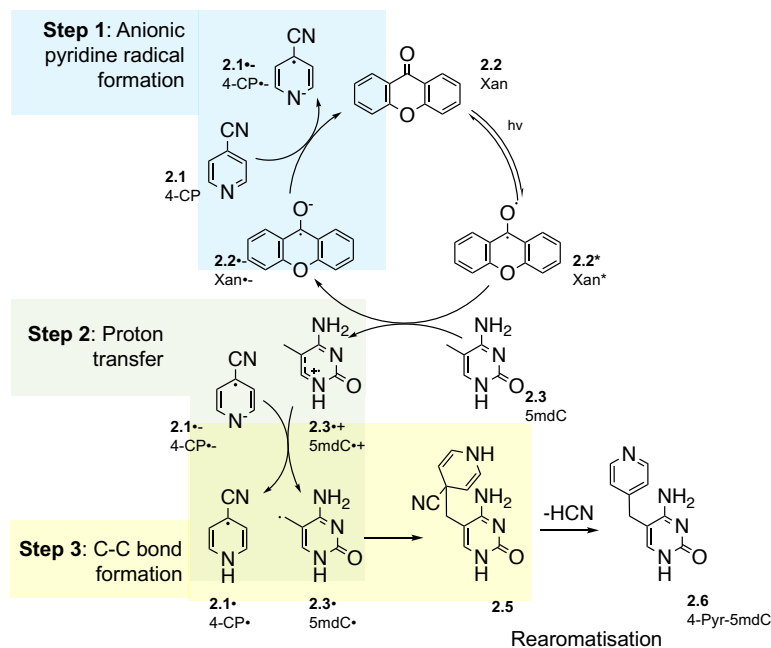
1. None of the RMSD values in the array is less than 1.0.
2. Conformers with an average RMSD value ranked within the top 5% are considered as the structural outliers.

DFT calculations were performed with Gaussian 16.¹⁰⁴ Structural optimisations and frequency calculations were conducted with the B3LYP-D3 functional and the 6-31G(d) basis set.^{105–108} Singlet-point energy values of optimised structures were calculated at the ω B97X-D/6-311++G(d,p) level of theory.¹⁰⁹

The deoxyribose ring in 5mdC ('dRib') is simplified to '-H' in the structures when running the calculations as the ring is far away from the site of the reaction.

2.3 Mechanistic revision

A. Initial Mechanism



B. Revised Mechanism

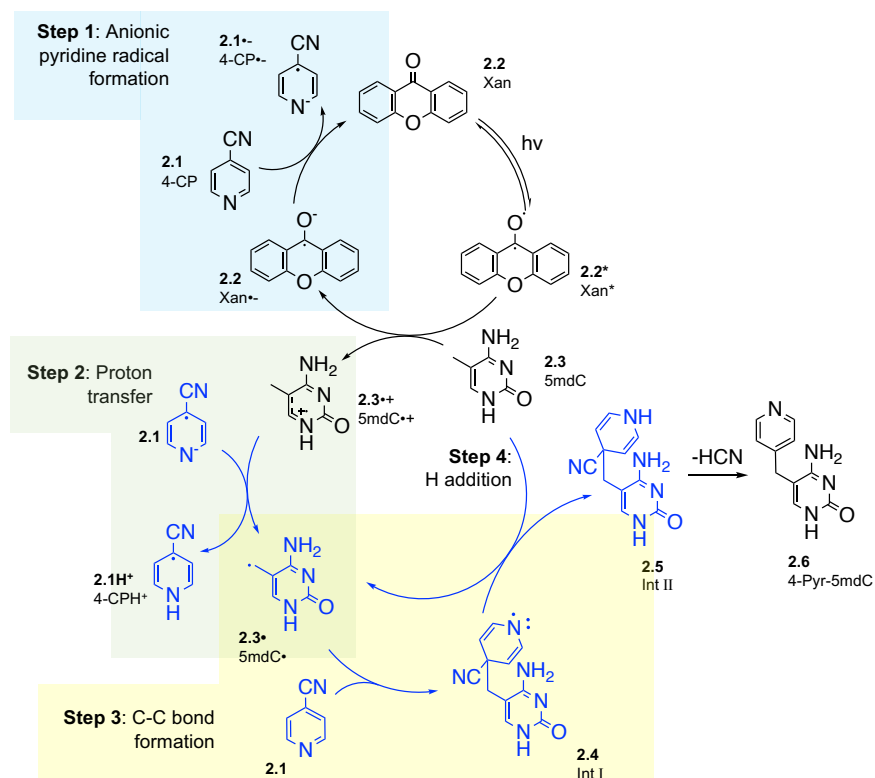


Figure 2.2. The initial and refined mechanism. A) The initial mechanism is based on previous literature.^{110–112} B) The refined mechanism upon computational studies. The changes from the initial proposal are coloured in blue. Here, the anionic cyanopyridine radical from step 1 is no longer essential for the sequential steps.

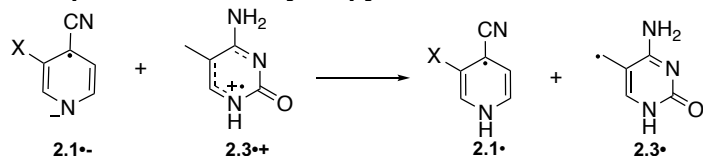
An initial mechanism has been proposed based on similar reactions in the literature (Figure 2.2.A).^{110–112} The literature suggests a radical-radical mechanism for the 5mdC-methyl radical formation (Step 2) and the carbon-carbon bond forming step (Step 3). Calculations were performed to verify and improve the initial proposal by exploring various potential pathways.

The refined proposal (Figure 2.2.B) involves an alternative mechanism from step 2 onwards. The triplet excited state xanthone, Xan^{T1} , abstracts an electron from 5mdC, producing the xanthone radical anion $\text{Xan}^{\bullet-}$ and the 5mdC radical cation $5\text{mdC}^{\bullet+}$. The $\text{Xan}^{\bullet-}$ transfer an electron to 4-cynopyridine, 4-CP, giving radical anion $4\text{-CP}^{\bullet-}$ and re-generating the ground state xanthone (Step 1). The proton transfer from $5\text{mdC}^{\bullet+}$ to 4-CP produces the neutral $5\text{mdC}^{\bullet}$ radical (Step 2). C-C bond occurs between a neutral pyridine and $5\text{mdC}^{\bullet}$ (Step 3). Int II is generated via hydrogen-atom abstraction from a neutral 5mdC molecule to the adduct Int I (Step 4). The final pyridination product, 4-Pyr-5mdC, comes from re-aromatisation and elimination of HCN from Int II.

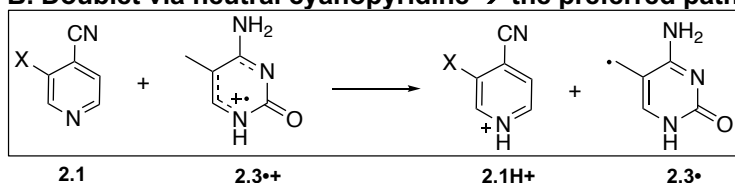
2.3.1 Step 2: Proton transfer

Various potential pathways have been considered for step 2 of the overall process. The neutral 5mdC radical may be generated via proton transfer from cationic 5mdC radicals to anionic cyanopyridine radicals, neutral cyanopyridine substrates or anionic xanthone radicals (Figure 2.3.A, B and C). The solvent, water, can also be involved in the proton transfer process (Figure 2.3.D). Thorough thermodynamic and kinetic studies are conducted on these potential pathways.

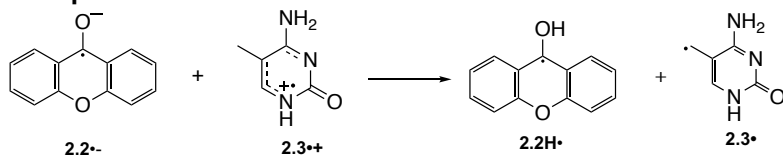
A. Triplet via anionic cyanopyridine radical



B. Doublet via neutral cyanopyridine → the preferred pathway



C. Triplet via anionic xanthone radical



D. Doublet via water

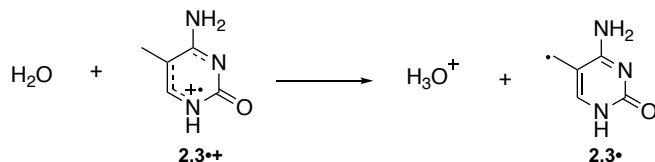


Figure 2.3. The possible pathways for the deprotonation of cationic 5mdC radical (step 2).

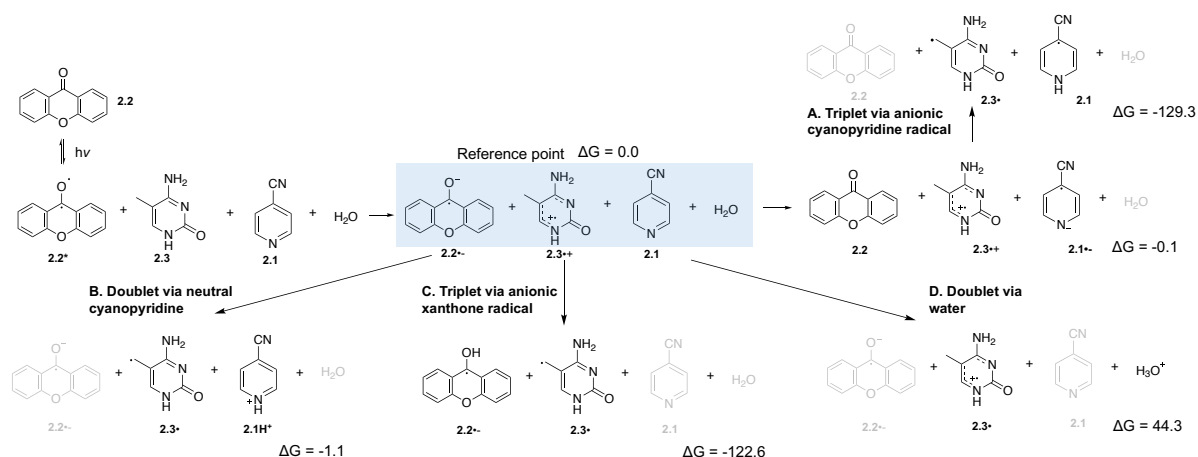


Figure 2.4. The thermodynamics of the possible step 2 pathways. Here, the ΔG calculations are based on 4-cyanopyridine. For clarity, structures that stay the same in the process are coloured in grey.

Thermodynamically (Figure 2.4), the triplet-based deprotonation pathways via anionic cyanopyridine derivatives (Figure 2.3.A) and anionic xanthone radical (Figure 2.3.C) are spontaneous. These values are large because the reactants are charged, and they come together to form neutral products. The doublet pathway via neutral cyanopyridine is spontaneous for most cyanopyridine substrates (Figure 2.3.B), except for those with an electron-withdrawing group at position 3 ($\Delta G = 1.0 - 12.2 \text{ kcal mol}^{-1}$). The deprotonation of cationic 5mdC radicals via water is thermodynamically unfavourable (Figure 2.3.D).

In the kinetic studies, two different mechanisms have been considered. The cationic 5mdC radical deprotonation can occur directly at the methyl group (mechanism C) or at the amine group (mechanism A) followed by a sequential intramolecular H transfer. Kinetically, triplet proton transfer via an anionic xanthone radical (Figure 2.3.C, $\Delta G^\ddagger(C) = 1.9 \text{ kcal mol}^{-1}$, $\Delta G^\ddagger(A) = 6.9 \text{ kcal mol}^{-1}$)^a is the most favourable, followed by the doublet pathway via neutral cyanopyridine substrates (Figure 2.3.B and 2.5, $\Delta G^\ddagger(C) = 14.1 - 36.7 \text{ kcal mol}^{-1}$, $\Delta G^\ddagger(A) = 5.6 - 49.3 \text{ kcal mol}^{-1}$) and the triplet pathway via anionic cyanopyridine radicals (Figure 2.3.A and 2.6, $\Delta G^\ddagger(C) = 12.9 - 28.5 \text{ kcal mol}^{-1}$, $\Delta G^\ddagger(A) = 8.6 - 28.2 \text{ kcal mol}^{-1}$).

^a The $\Delta G^\ddagger$ in the text are from calculations at the $\omega B97X-D/6-311++G(d,p)/\omega B97X-D/6-31G(d)$ level of theory. At the $\omega B97X-D/6-311++G(d,p)/B3LYP-D3/6-31G(d)$ level of theory, $\Delta G^\ddagger = -1.13 \text{ kcal mol}^{-1}$.

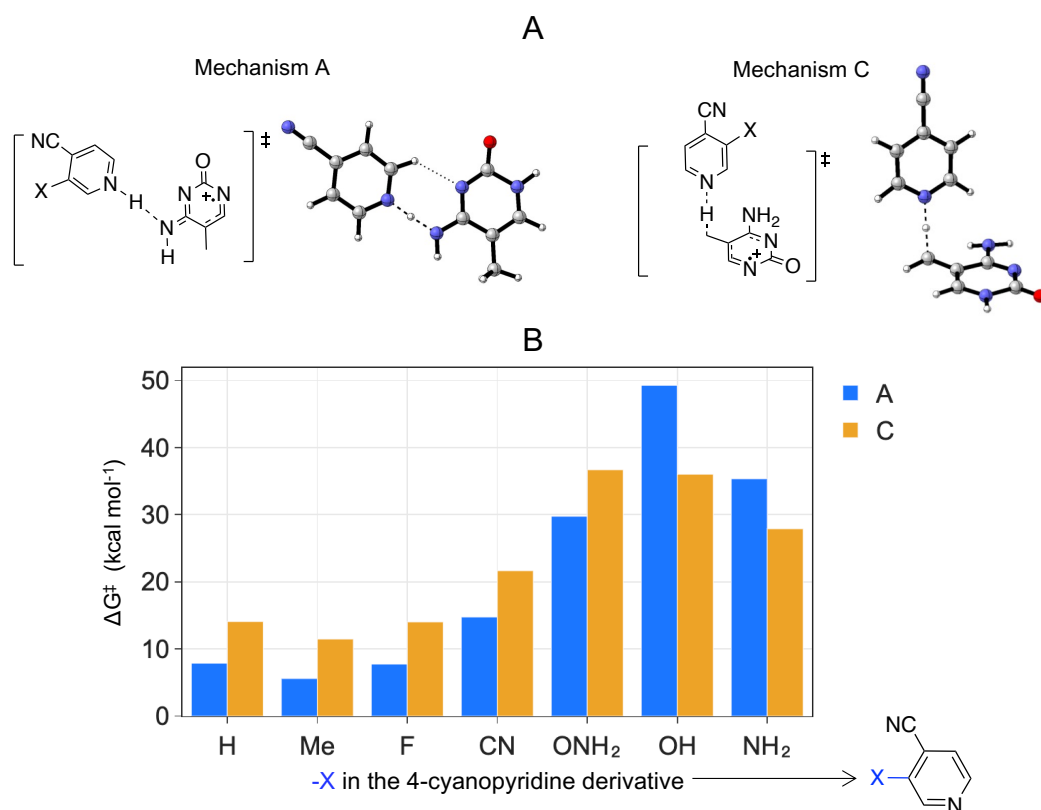


Figure 2.5. Kinetic studies of the doublet pathway via neutral cyanopyridine substrates, *ie* the reaction in Figure 2.3.B. A) The two possible TSs for the proton transfer between neutral cyanopyridine substrates and cationic 5mdC radicals are given. 3D structures of the TSs were given for systems with X = H. B) The bar chart compares the $\Delta G^\ddagger$ of the two possible TSs for each cyanopyridine derivative.

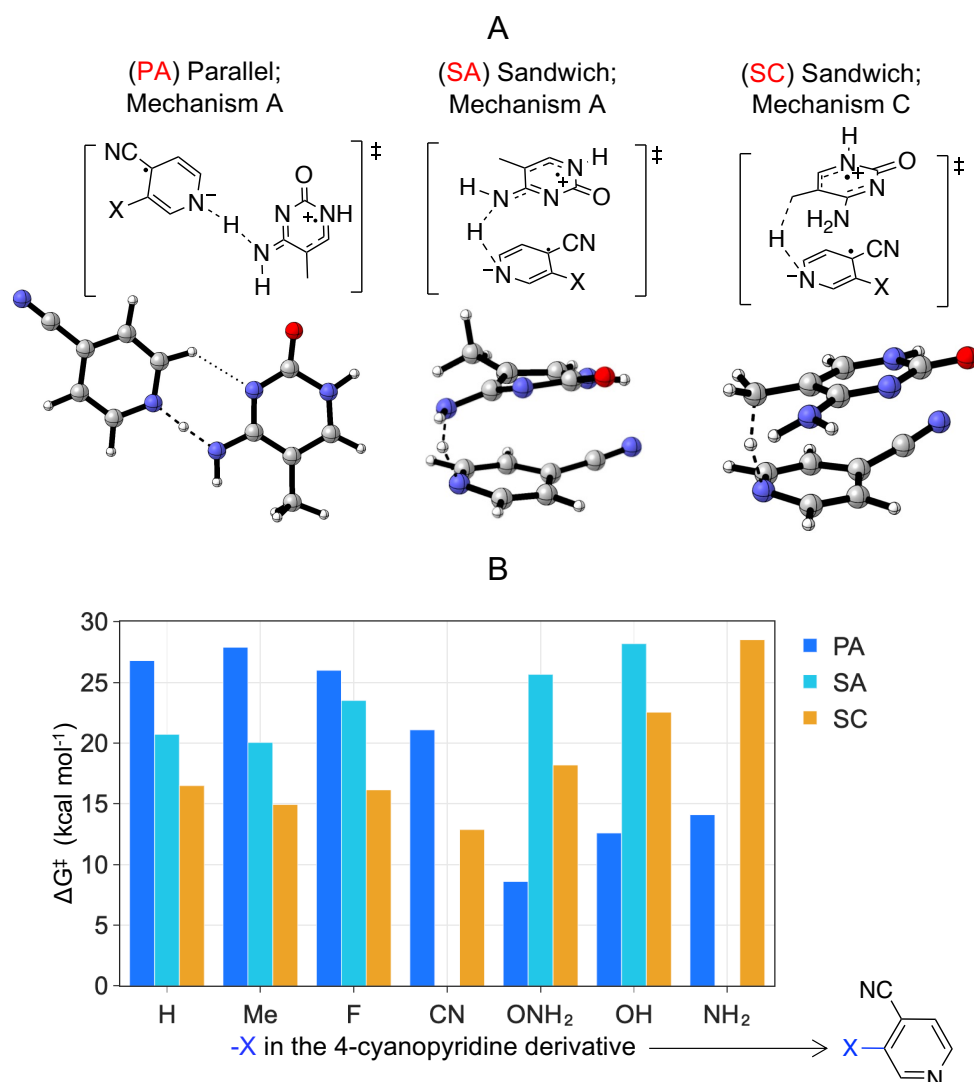


Figure 2.6. Kinetic studies of the triplet pathway via anionic cyanopyridine radical, *ie* the reaction in Figure 2.3.A. A) The three possible TSs for the proton transfer between anionic 4-cyanopyridine radicals and cationic 5mdC radicals are given. In the process via mechanism A, the ring systems can take up either a sandwich or parallel conformation. 3D structures of the TSs were given for systems with X = H. Mechanism C with a parallel conformation is presumably high in energy and such conformation has not been found in conformational searches. B) The bar chart compares the $\Delta G^\ddagger$ of the three possible TSs for each cyanopyridine derivative. A few of the SA TSs have not been localised. In these cases, conformation changes from the sandwich conformation to the parallel conformation always occurred during bond scan calculations.

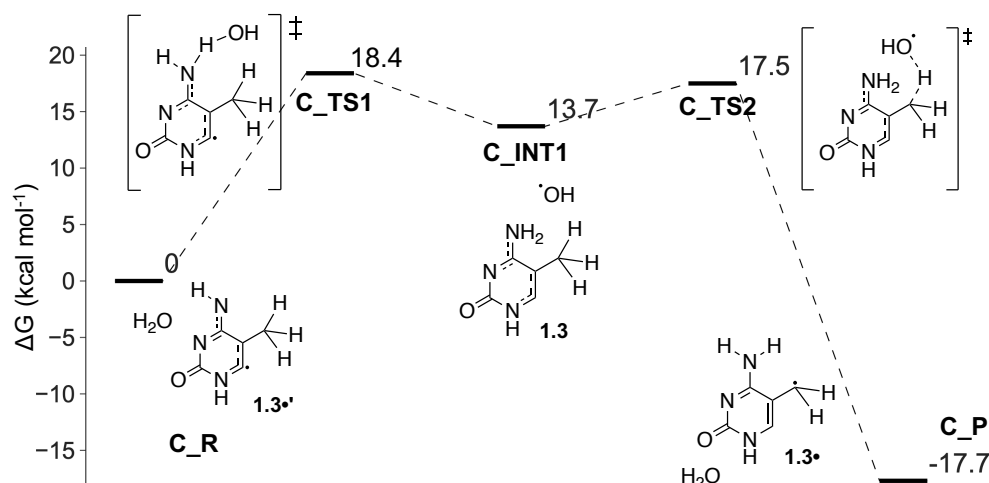


Figure 2.7. An energy profile for the intramolecular H transfer of 5mdC via water.

With mechanism A, a sequential intramolecular H transfer is required to generate 5mdC•. The H transfer can occur directly or with the assistance of a water molecule as a catalyst. Calculations show that the 5mdC intramolecular H transfer via water ($\Delta G^\ddagger = 18.4 \text{ kcal mol}^{-1}$) is the lower energy pathway compared to without water ($\Delta G^\ddagger = 25.7 \text{ kcal mol}^{-1}$) (Figure 2.7).

Based on the above calculations, the two triplet pathways (Figure 2.3.A and C) and the doublet pathway with neutral cyanopyridine as the base (Figure 2.3.B) are all theoretically possible to occur. However, triplet pathways involve the collision between two radical species. Due to the excess of cyanopyridine (300 mM) compared to xanthone (10 mM, *ie* catalytic amount) used in the reaction, the concentration of anionic radical species in the reaction will be low compared to neutral cyanopyridine.⁹⁸ The chance of a collision between two radicals in the triplet pathways is low. Furthermore, the triplet pathway via anionic xanthone radical (Figure 2.3.C) does not support the experimental results. In this pathway, cyanopyridine substrates are not required to accept net charges and changing the electronic nature of the cyanopyridine ring should not affect the percentage yield of the reaction to a large extent. The pathway via neutral cyanopyridine derivatives (Figure 2.3.B) supports the experimental data and explains the variation in the percentage yield of the cyanopyridine substrates (see Chapter 2.4. for more details). Therefore, 5mdC•+ is most likely to be deprotonated by neutral cyanopyridine to form 5mdC•.

For the doublet pathway via neutral cyanopyridine derivatives (Figure 2.3.B), direct deprotonation at the methyl group (mechanism C) is more likely to occur than deprotonation at the amine (mechanism A). The deprotonation at the amine group is less thermodynamically favourable compared to deprotonation at the methyl group (for example, for 4-cyanopyridine, $\Delta G(C) = -1.1$ and $\Delta G(A) = 14.6 \text{ kcal mol}^{-1}$). Additionally, the sequential intramolecular H transfer in mechanism A has a high kinetic barrier of $18.4 \text{ kcal mol}^{-1}$. Thus, mechanism C is more favourable than mechanism A.

2.3.2 Step 3: C-C bond formation

Step 3 of the reaction involves a C-C bond formation between C4 in cyanopyridine and -CH₂ in a 5mdC radical (Figure 2.8). Radical-radical coupling mechanisms have been proposed for similar reactions in literature. The coupling can occur between a neutral cyanopyridine radical and a 5mdC radical via a neutral triplet TS (Figure 2.8.A).¹¹⁰ As neutral cyanopyridine radicals can only be generated in the minor triplet pathway of step 2, the kinetics of other possible pathways were examined. An anion triplet pathway is also plausible, where the coupling happens between an anionic cyanopyridine radical and a neutral 5mdC radical (Figure 2.8.B).^{113–115} Besides radical-radical coupling terminations, the C-C bond formation can also take place via a doublet or a singlet TS. In the doublet pathway, the C-C bond is formed between a neutral cyanopyridine and a 5mdC radical (Figure 2.8.C). In the singlet pathway, the radicals convert into ions and the C-C bond formation occurs between ions (Figure 2.8.D).

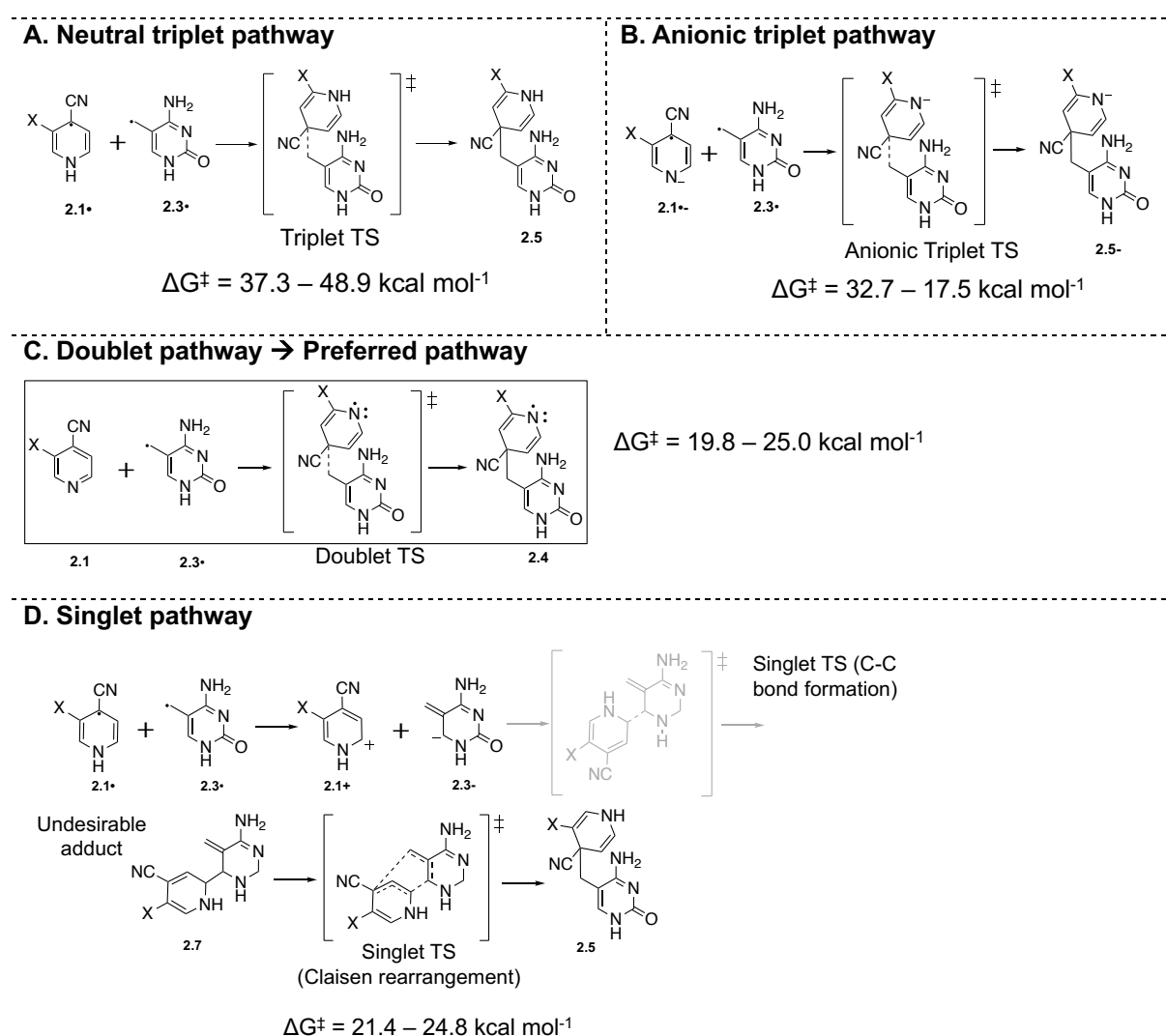


Figure 2.8. The possible C-C bond formation pathways. The ΔG and $\Delta G^\ddagger$ values are given as a range as pyridine substrates are involved in the process.

Transition state analyses were conducted on the four possible pathways. Neutral triplet TSs, anionic triplet TSs and doublet TSs were found. $\Delta G^\ddagger$ values of these TSs were calculated (Figure 2.9). Bond scan calculations of the adduct with a singlet state setting reveals that the singlet pathway is a two-step process (Figure 2.8.D). The ion-ion coupling reaction leads to an undesirable adduct, which can undergo a Claisen rearrangement reaction to give the desirable adduct. The TS of the ion-ion coupling reaction was not obtained, presumably due to the barrierless nature of the process. The Claisen rearrangement reaction TSs were found and optimised. $\Delta G^\ddagger$ values of neutral triplet TSs are significantly higher compared to the other pathways. Therefore, a radical-radical coupling mechanism via a neutral triplet TS is not the favourable pathway for the C-C bond formation. The difference in $\Delta G^\ddagger$ between the three other possible TSs is often less than 2 kcal mol⁻¹. In theory, the three possible pathways are all likely to occur. As explained in the last section (Chapter 2.3.1 Step 2: Proton Transfer), the concentration of radical species is expected to be low relative to the concentration of neutral cyanopyridine. Therefore, C-C bond formation is most likely to occur via the doublet pathway.

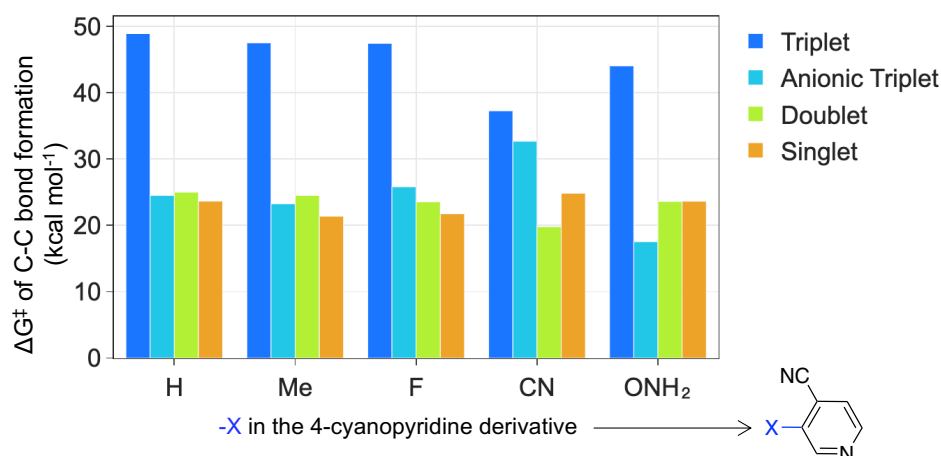
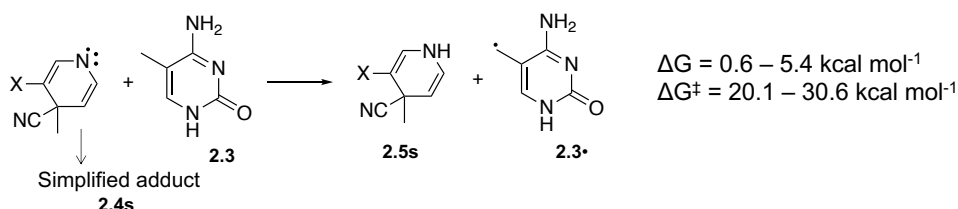


Figure 2.9. The kinetic study of the four possible C-C bond formation pathways: The bar chart compares the $\Delta G^\ddagger$ of the TSs from the four possible pathways for each cyanopyridine derivative. -X is the substituent at position C3 of the pyridine.

2.3.3 Step 4: Hydrogen atom addition

Potential pathways for step 4 were deduced after establishing the mechanism for step 2 and 3. Step 4 should be a H addition reaction of the doublet adduct, Int I. The H addition can occur either via a neutral 5mdC molecule or via the solvent of the reaction, *ie* water. The structure of Int I was simplified for the computational investigation.

A. H addition via 5mdC



B. H addition via water

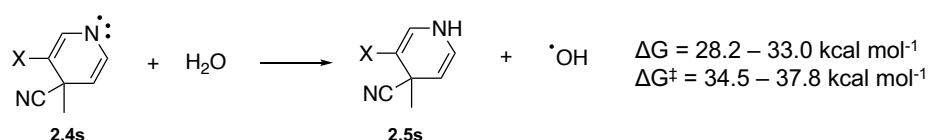


Figure 2.10. The possible H addition pathways. The ΔG and $\Delta G^\ddagger$ values are given as a range as pyridine motif-containing adducts are involved in the process.

The H addition via 5mdC is more favourable thermodynamically than the pathway via water (Figure 2.10). Kinetically, the pathway via water has a considerably higher activation barrier compared to the pathway via 5mdC (Figure 2.11). H addition via 5mdC is more preferable compared to via water. However, water is the solvent and present in significantly larger quantities compared to 5mdC. It is highly probable that the pathway via water will still occur.

Various possible TS conformations have been considered for the H addition via 5mdC. The reaction can occur with a parallel or a sandwich conformation via direct H transfer from the methyl group (Mechanism C) or H transfer from the amine group followed by a 5mdC intramolecular H transfer (Mechanism A). There are four possible TS conformations: SA (sandwich, mechanism A), PA (parallel, mechanism A), SC (sandwich, mechanism C) and PC (parallel, mechanism C). Among the possible TS conformations for pathways via 5mdC, TSs with a SC conformation have the lowest $\Delta G^\ddagger$ value across a range of cyanopyridine substrates. The $\Delta G^\ddagger$ difference between the four possible TSs for pathways via 5mdC is relatively small. Therefore, mechanism A may still be competitive. Taking 3-methyl-4-cyanopyridine as an example, the four TSs have a $\Delta G^\ddagger$ difference within $2.5 \text{ kcal mol}^{-1}$.

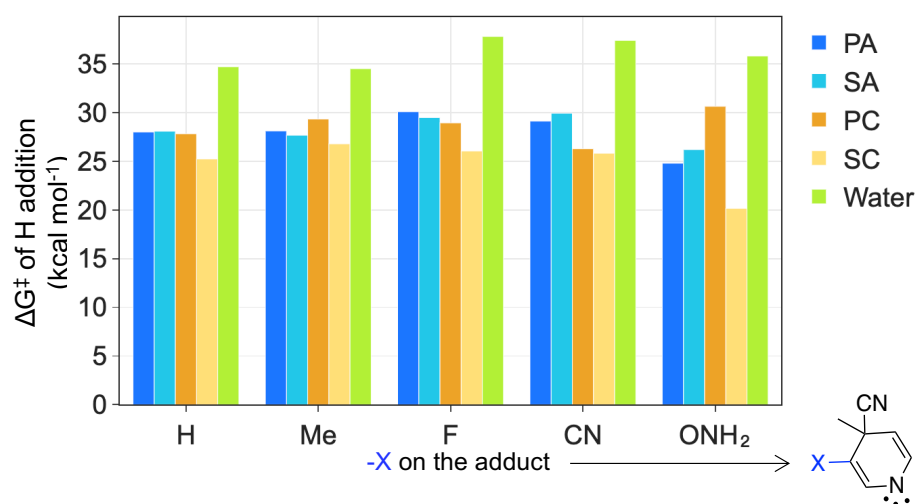
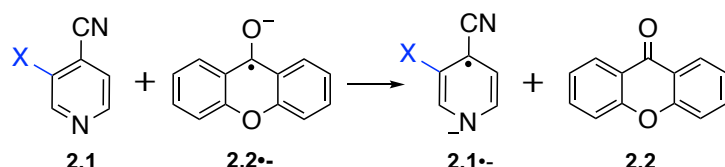


Figure 2.11. Kinetic studies of the H addition process: the bar chart compares $\Delta G^\ddagger$ values of possible H addition processes via water and via 5mdC.

2.4 Variation in percentage yield

Step 1 and 2 of the reaction (Figure 2.12) involve a charge transfer to the pyridine substrate. Changing the substituent at position C3 varies the electronic properties of the pyridine substrates, which impacts the outcome of step 1 and 2. Thorough thermodynamic and kinetic studies have been conducted to explain the variation in percentage yield between pyridine substrates.

A. Step 1: Electron transfer



B. Step 2: Proton transfer

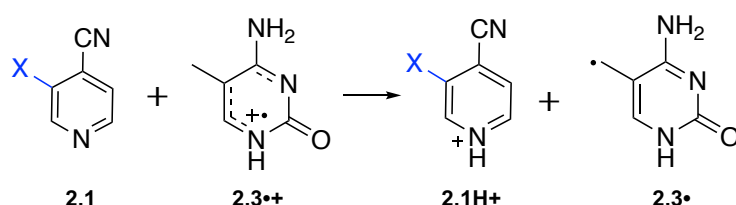


Figure 2.12. The focus of the investigation on understanding the variation in percentage yield – step 1 and 2 of the reaction in the revised mechanism (Figure 2.2.B)

2.4.1 Thermodynamic study

In step 1, the pyridine needs to accept an electron. Substrates with an electron-donating group are expected to have a relatively negative ring and are predicted to be unfavourable thermodynamically. Step 2 is a proton transfer process from 5mdC to the pyridine. Substrates with an electron-withdrawing group are predicted to be unfavourable thermodynamically due to a relatively positive ring. The nature of step 1 and 2 implies competing impacts on the yield of the reaction.

ΔG values of step 1 and 2 (ΔG_{step1} and ΔG_{step2}) were computed for a range of 4-cyanopyridine substrates. The correlation between the thermodynamic result and the electronic nature of the cyanopyridine rings was explored through comparisons with Hirshfeld charges. Specifically, the sum of Hirshfeld charges on the substituent of neutral cyanopyridine substrates was calculated and analysed. Besides the calculated datasets, Hammett sigma constants from the literature were also considered for validation.¹¹⁶ σ_x comes from experiments and is based on the ionisation process of benzoic acid (eq. 1):

$$\sigma_x = \log K_x - \log K_H \quad (1)$$

K_H is the ionisation constant for benzoic acid in water. K_x is the ionisation constant for a meta-substituted benzoic acid. σ_x is a collective measure of the substituent's ability to withdraw or donate electrons from the site of reaction. The σ_x of an electron-withdrawing substituent, such as -CN, tends to be positive and vice versa for the σ_x of an electron-donating substituent.

ΔG_{step1} values are plotted against σ_x in Figure 2.13.A. A noticeable change in trend is observed at approximately $\sigma_x = 0.2$. When $\sigma_x > 0.2$, the correlation between ΔG_{step1} values and Hammett sigma constants follows a negative linear relationship. Electron-withdrawing groups withdraw charges from the ring system and make the ring more positive, favouring the electron transfer process. Hence, substrates with a strong electron-withdrawing substituent have a more negative ΔG_{step1} than substrates with a weak electron-withdrawing or an electron-donating substituent. When $\sigma_x < 0.2$, ΔG_{step1} values stay the same regardless of the electron-donating strength of the substituent.

Charge analyses provide insights into the electronic nature of the cyanopyridine rings. The ΔG_{step1} vs the sum of Hirshfeld charges on the C3 substituent plot in Figure 2.13.B gives a rough positive linear relationship (Pearson's $r = 0.75$). Substrates with a stronger electron-withdrawing substituent have negative Hirshfeld charges on the C3 substituent and a negative ΔG_{step1} value. On the other hand, the charges located on the electron-donating substituents are very similar across different cyanopyridine substrates. Similarly, the variation of ΔG_{step1} values for substrates with an electron-donating group is insignificant.

For step 2, a positive linear correlation is observed in the ΔG_{step2} vs σ_x plot (Figure 2.13.C), which aligned with the expectation. Electron-withdrawing groups induce a relatively positive ring system, making proton acceptance less favourable and contributing to the positive ΔG_{step2} . Conversely, electron-donating groups donate electron density to the ring and make the process more favourable thermodynamically.

The plot of ΔG_{step2} vs the sum of Hirshfeld charges on the substituent demonstrates a negative linear trend (Pearson's $r = -0.88$, Figure 2.13.D), which confirms the link between the thermodynamics of step 2 and the electronic nature of the cyanopyridine rings. The sum of Hirshfeld charges on the substituent is negative, implying a positive ring for substrates with an electron-withdrawing group. This leads to the positive ΔG_{step2} and the low percentage yield. The opposite applies for substrates with an electron-donating group.

Based on the above analyses, two conclusions can be drawn:

1. The outcome from the thermodynamics cannot explain the low percentage yield for substrate with an electron-donating group. In step 1, irrespective of the strength of the electron-donating ability of the substituent on the substrate, comparable ΔG_{step2} and charge analysis outcomes are observed when compared to high-yielding substrates, such as 4-cyanopyridine. Therefore, other factors, such as reaction kinetics, must be considered.
2. Substrates with an electron-withdrawing group ($X = \text{CN}, \text{SCN}, \text{F}$) are unfavoured thermodynamically in step 2, which matches with the prediction. The positive ΔG_{step2} contributes to the low percentage yield of the pyridination reaction.

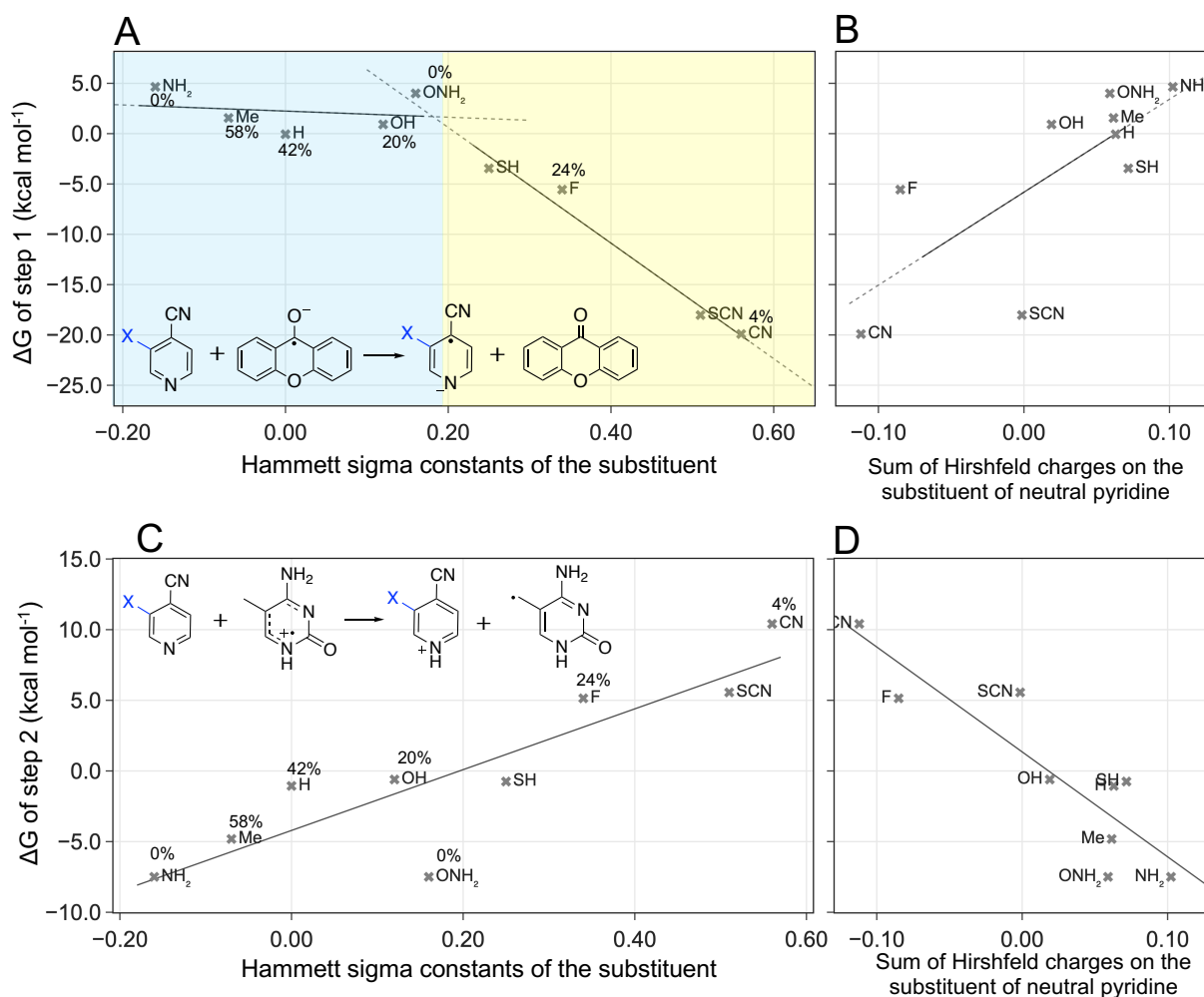


Figure 2.13. An overview of the thermodynamic and charge analysis results for step 1 and 2: A) ΔG_{step1} vs Hammett sigma constants of the substituent plot (Gradient of the best-fit lines: blue region: -3.0; yellow region: -57.4). The change in trend occurs at approximately $\sigma_x = 0.2$, which is indicated by the background colouring in plot A. B) ΔG_{step1} vs the sum of Hirshfeld charges on the substituent of the neutral cyanopyridine substrate plot. C) ΔG_{step2} vs Hammett sigma constants of the substituent plot (Gradient of the best-fit lines: 21.5). D) ΔG_{step2} vs the sum of Hirshfeld charges on the substituent of neutral cyanopyridine substrates plot. Each point on the plots represents a cyanopyridine derivative. The substituent labels are next to the data points. The chemical equations of step 1 and 2 for pyridination with 4-cyanopyridine are drawn and placed inside plot A and C, respectively. The percentage yield values are given next to the data points in plot A and C.

2.4.2 Kinetic study

The low percentage yield observed in reactions involving substrates with an electron-donating group can be explained by looking into the kinetics of step 2. In Chapter 2.3.1. Step 2: Proton Transfer, the major pathway for the deprotonation of cationic 5mdC radical is confirmed to be via neutral cyanopyridine substrates. Mechanism C is preferred with deprotonation at the methyl group.

The $\Delta G^\ddagger$ values for deprotonation are much higher for cyanopyridine substrates with -OH, -NH₂ and -ONH₂ groups ($\Delta G^\ddagger = 27.9 - 36.7$ kcal mol⁻¹) compared to other cyanopyridine substrates ($\Delta G^\ddagger = 11.0 - 22.9$ kcal mol⁻¹) (Figure 2.4). Mechanism C TSs for the doublet pathway with different cyanopyridine substrates have very similar structures (Figure 2.14.A). Therefore, the high kinetic barrier is due to the initial reactant dimer conformation between neutral cyanopyridine and 5mdC•+. The lowest energy reactant dimer conformation with the -NH₂ substituted cyanopyridine is stabilised by H-bonding (Figure 2.14.B). The stabilising H bond is not presented in the TS structure. The cleavage of the stabilising H bond is required to arrive at the TS. Thus, the energy difference between the reactant dimer and the TS is more significant for substrates with H-bond donor (X = -NH₂, -ONH₂ and -OH).

In summary, cyanopyridine substrates with a H-bond donor group, which also happen to be electron-donating, are kinetically unfavourable in step 2 of the overall reaction.

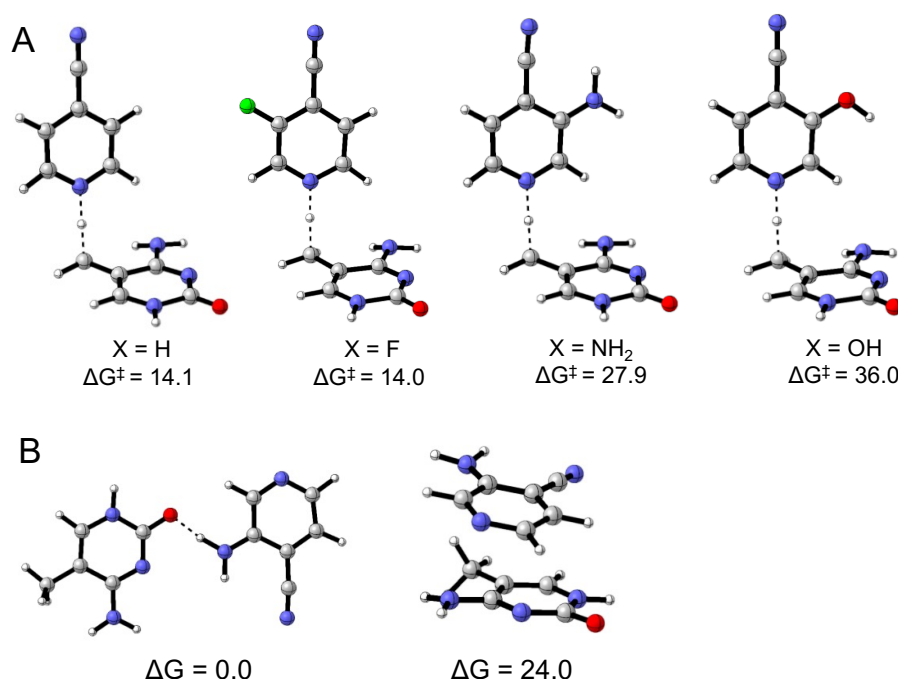


Figure 2.14. Explaining the high kinetic barriers for substrates with a H-bond donor group in step 2. A) Mechanism C TS structures with different cyanopyridine substrates. B) Conformers of the cationic 5mdC radical and 3-amino-4cyanopyridine substrate

2.5 Summary

This chapter focuses on understanding the reactivity of 5-methylcytosine pyridination reactions with the established computational methods. A combination of MM and DFT calculations were utilised to explore the PES and potential pathways.

The reaction mechanism for the pyridination of 5-methylcytosine was revised based on calculations about possible competing pathways. As opposed to the radical-radical mechanism from the literature,^{110–112} this study suggests that the generation of neutral 5mdC• radical (step 2) and the C-C bond formation (step 3) are via neutral pyridine substrates. The calculations support the kinetic feasibility of the revised mechanism compared to the original proposal (Figure 2.2.B). An overall energy diagram is presented in SI Figure S2.1. Experimentally, the chance of a collision between two radicals is also low as a catalytic amount of xanthone was used.

The variation in the percentage yield can be explained by looking at step 2 of the reaction, *ie* the proton transfer from 5mdC•+ radicals to neutral pyridine substrates. Substrates with an electron-withdrawing group are thermodynamically unfavourable in step 2. Substrates with an electron-donating group are unfavourable kinetically. The electron-donating substituents also happen to be good H-bond donors, which leads to the high kinetic barrier in step 2.

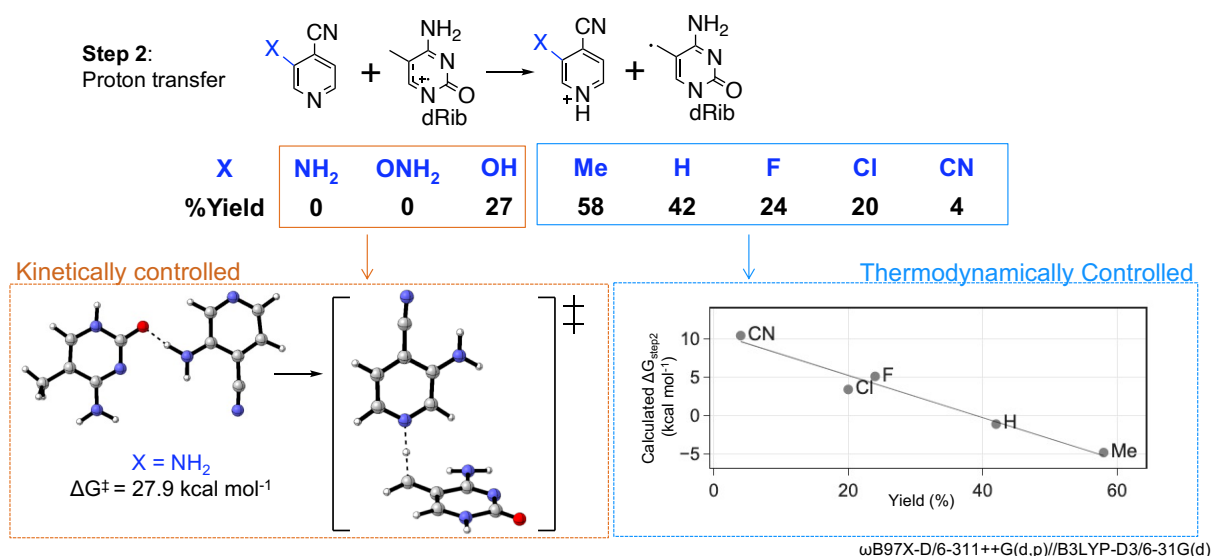


Figure 2.15. Summary: understanding the variation in the percentage yield

Chapter 3

Reactions with diarylprolinol silyl ether catalyst: Investigating large and structurally flexible systems

3.1 Background

In the reactivity investigation of large and structurally flexible organic systems, thorough exploration of the conformational space is a challenge. Additional thought needs to be put into innovating the computational workflow for studying complex systems efficiently and effectively.

This chapter focuses on complex organic systems with a large conformational space, specifically reactions involving diarylprolinol silyl ether catalysts from Melchiorre *et al.*¹¹⁷ Recently, they have proposed a photocatalytic route to synthesise 1,4-dicarbonyl compounds with stereochemical control using 1,4-dihydropyridine (DHP) and diarylprolinol silyl ether catalysts (the ‘amine catalysts’ below, Figure 3.1). In the optimised trial with catalyst C, a high yield of 88% has been achieved with a reasonable enantiomeric excess (*ee*) of 76%.

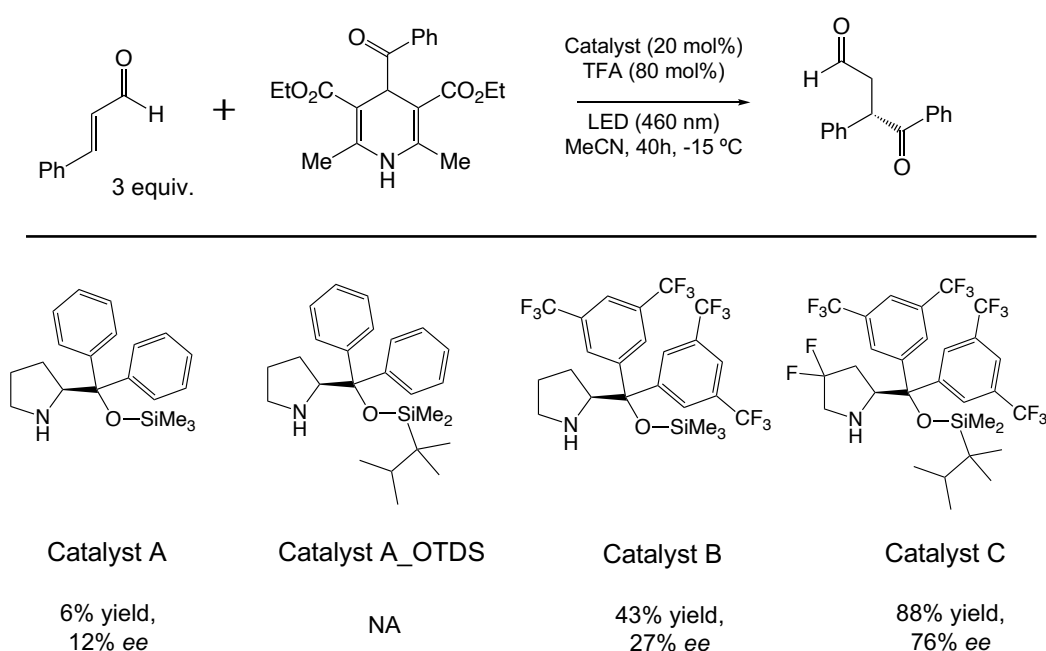


Figure 3.1. Enantioselective synthesis of 1,4-dicarbonyls with amine catalysts from the work of Melchiorre *et al.*^{117,118} The experimental result of the reaction with catalyst A comes from Yu *et al.*,¹¹⁸ where Ru(I) complex was used for photoinitiations. (OTDS = oxy-thexyldimethylsilyl group)

Diarylprolinol silyl ether systems were first proposed by Jørgensen *et al.* to achieve enantioselective sulfenylation of aldehydes in 2005.^{119,120} Further research efforts in the last decade have extended the

application of the catalytic systems to other functionalisation reactions at positions neighbouring the carbonyl group via the formation of either an iminium or enamine molecule.¹¹⁹ The catalysis of diarylprolinol silyl ether compounds has previously been studied with computational methodologies on functionalisation reactions with ionic pathways.¹²¹ In 2014, Seebach *et al.*¹²² performed density functional theory (DFT) calculations on various amine-catalysed reactions with ionic mechanisms and derived trends that correlate the size of the substituent with the reaction enantioselectivity.

The applications of diarylprolinol silyl ethers in radical pathways have not been widely explored. Unlike in ionic approaches, acyl radicals are highly reactive and achieving good enantioselectivity becomes more challenging via a radical route.¹²³ Following the success in 2013,¹²⁴ Melchiorre *et al.* have reported several enantioselective photochemical reactions with the use of amine catalysts, which include the 1,4-dicarbonyl synthesis with DHP.^{117,125–127} However, the origin of the enantioselectivity for these reactions remains unexplored. In 2021, Duarte *et al.*¹²⁸ conducted computational studies on the selectivity of α -chiralbicyclo[1.1.1]pentane reactions. The reaction required amine-based catalysts and followed a similar photochemical radical route. Due to the nature of the functional groups attached to the carbonyl, the selectivity was mainly controlled by local non-covalent interactions (NCIs). Good experimental enantioselectivity has also been reported on reactions with substrates that are relatively inert to form strong NCIs, which includes the studies from Melchiorre *et al.*^{117,129}

In this chapter, the photocatalytic 1,4-dicarbonyl synthesis from Melchiorre *et al.* was investigated computationally. The enantioselectivity of the reactions is explained from the perspective of conformational changes and structural deformations. A computational procedure is proposed for investigating large and structurally flexible chemical systems. Automated Python pipelines have been written and applied to facilitate this effective and efficient procedure, which is potentially transferable to related chemical systems.

3.2 Computational methods

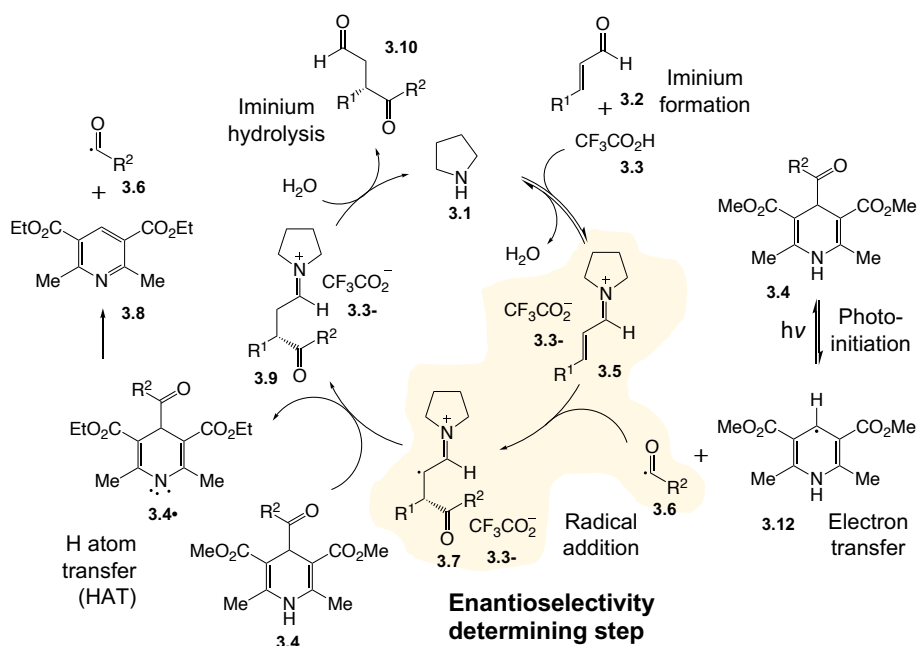
Conformational searching calculations were conducted in MacroModel (v11.7)³⁹ with MMFF¹⁵ and a setting of 2000-4000 steps as the maximum number of steps. Conformers within an energy window of 41.6 kJ mol⁻¹ (ie an equivalent of 10 kcal mol⁻¹) were saved for further analyses.

DFT optimisations were performed with Gaussian 16¹⁰⁴ at the B3LYP/6-31G(d) level of theory.^{105–107} Single-point energy calculations were conducted at the ω B97X-D/6-311++G(d,p) level of theory.¹⁰⁹ Inclusion of SMD solvent models¹³⁰ in single-point energy calculations was also considered during benchmarking of the transition state (TS) $\Delta G^\ddagger$.

A pipeline has been built to facilitate automation in computational investigations (https://github.com/Goodman-lab/Script_im). The scripts were written in Python 3 using Jupyter Notebook (6.1.4). The used packages were as below: NumPy (1.19.2), Pandas (1.1.3), Matplotlib (3.3.2), Collection (0.1.6), Biopython (1.78),¹³¹ Scikit-learn¹³² and GoodVibes¹³³.

3.3 Mechanism and the enantioselectivity-determining step

A. Mechanism 1 → the preferred mechanism



B. Mechanism 2

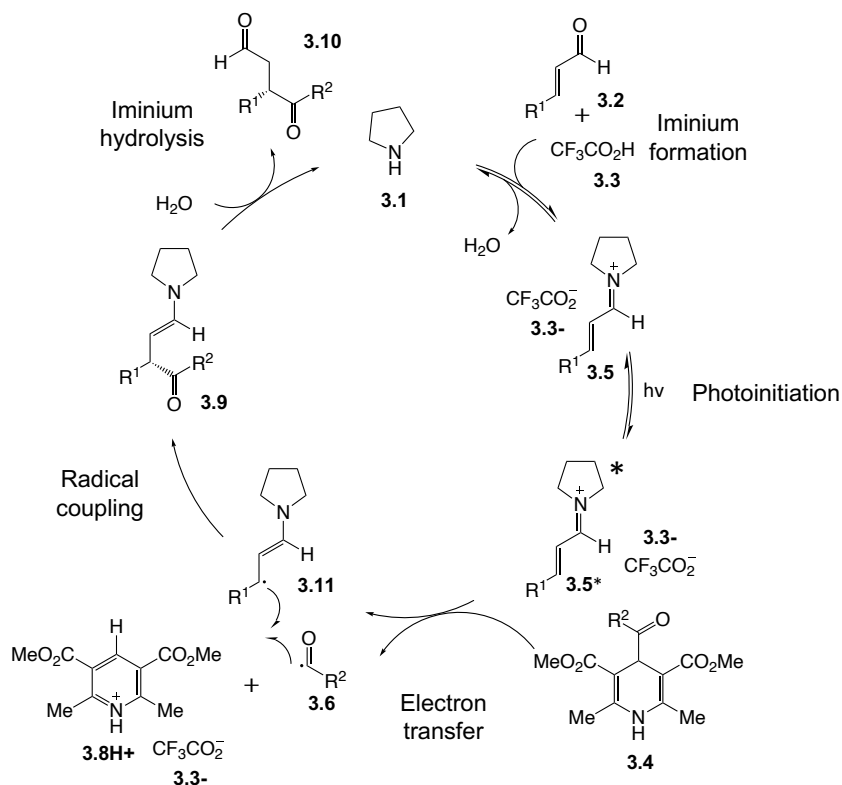


Figure 3.2. Possible mechanisms for the enantioselective synthesis of 1,4-dicarbonyls with amine catalysts. A) In mechanism 1,¹¹⁷ the C-C bond formation step is a radical addition process. B) In mechanism 2,^{134,135} the C-C bond formation step is a radical coupling process.

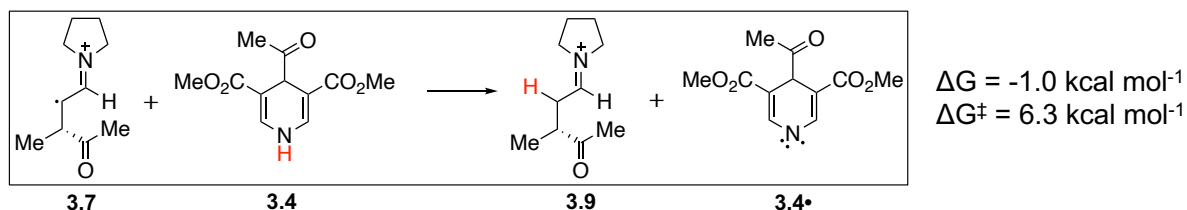
Melchiorre *et al.* have proposed the catalytic cycle (Figure 3.2.A, Mechanism 1) for the photocatalytic 1,4-dicarbonyl reaction.¹¹⁷ Amine catalysts form conjugated iminium ions with an enal and activate the β carbon of the iminium ion for incoming attacks.¹¹⁹ The DHP molecule absorbs a photon, which leads to homolytic fissions and the release of an acyl radical. The C-C bond formation between the iminium ion and the acyl radical occurs via a doublet pathway. The radical addition produces a cationic radical adduct as the product. The adduct needs to go through a H atom transfer (HAT) process before the iminium hydrolysis. Based on the possibility that the iminium ion could also be excited upon absorption of a photon, an alternative mechanism was considered (Figure 3.2.B, Mechanism 2).^{134,135} The excited iminium ion triggers an electron transfer process with the DHP, which leads to the generation of the acyl radical and the enamine radical. The two radical species can go through a coupling reaction to form the C-C bond via a triplet pathway.

The C-C bond formation step for mechanism 1 and 2 were considered computationally at the ω B97X-D/6-311++G(d,p)//B3LYP/6-31G(d) level of theory. A simplified system has been used for the mechanistic study (Figure 3.2). The activation energy for the radical addition process ($\Delta G^\ddagger = 4.7 \text{ kcal mol}^{-1}$) in mechanism 1 is much lower than the radical coupling process ($\Delta G^\ddagger = 24.6 \text{ kcal mol}^{-1}$) in mechanism 2, which suggests that the reaction proceeds via radical attack to the ground state iminium. Furthermore, the experiments were conducted with radiation at $\lambda_{\text{max}} = 460 \text{ nm}$, which is not optimal for exciting iminium ions. The optimal wavelength used for exciting conjugated iminium is 420 nm, as reported in the literature.^{134,135} Therefore, mechanism 1 is preferred over mechanism 2.

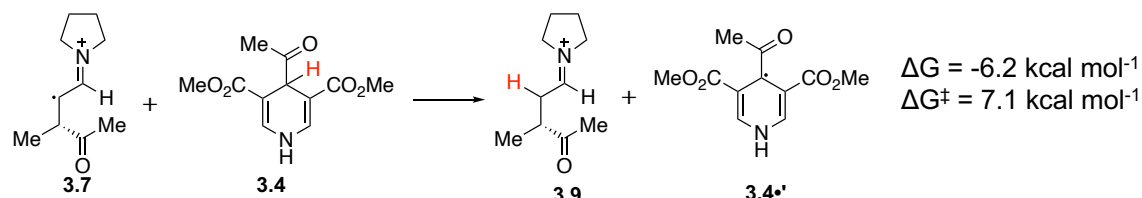
The sequential step of the radical addition process (*ie* the hydrogen atom transfer (HAT)) in mechanism 1 was investigated in terms of its feasibility and the mechanism of action. Three possible pathways were proposed based on results from conformational studies of the cationic radical adduct and DHP dimeric system (Figure 3.3). In pathway C1 and C2, H atoms from DHP are directly transferred to the site of carbon in the adduct. In pathway O, the H atom is transferred to the O in the carbonyl group and sequential rearrangements are required to arrive at the desired product. Pathway O is derived based on the strong H bond between the H-N in the DHP ring and the O in the carbonyl group of the adduct, which appears in many low-energy adduct-DHP dimers (Figure 3.4).

Pathway O is thermodynamically unfavourable. Optimisations of the product dimers from various distinct conformations always led to the reverse of the pathway O process. Optimised geometries of the reactants were obtained instead from the calculations. Therefore, pathway O can be eliminated from the list of possible pathways. Pathway C1 and C2 are competing processes and likely to coexist in the reaction flask. Both processes are exothermic and have a similar kinetic barrier. The $\Delta G^\ddagger$ of pathway C1 ($\Delta G^\ddagger = 6.3 \text{ kcal mol}^{-1}$) and C2 ($\Delta G^\ddagger = 7.1 \text{ kcal mol}^{-1}$) differ by less than 1 kcal mol^{-1} . Pathway C2 leads to the generation of a DHP radical species, which cannot release the carbonyl radical spontaneously. As DHP is the limiting reagent of the reaction, the HAT process might be a percentage yield-limiting step due to the existence of pathway C2.

A. Pathway C1 → the preferred pathway



B. Pathway C2



C. Pathway O

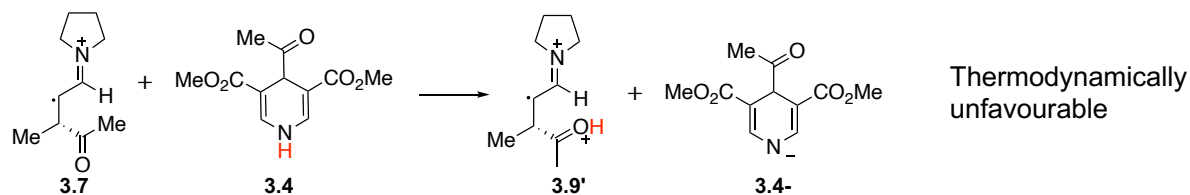


Figure 3.3. Three possible pathways of the H atom transfer (HAT) process. The energy values are reported at the ω B97X-D/6-311++G(d,p)//B3LYP/6-31G(d) level of theory.

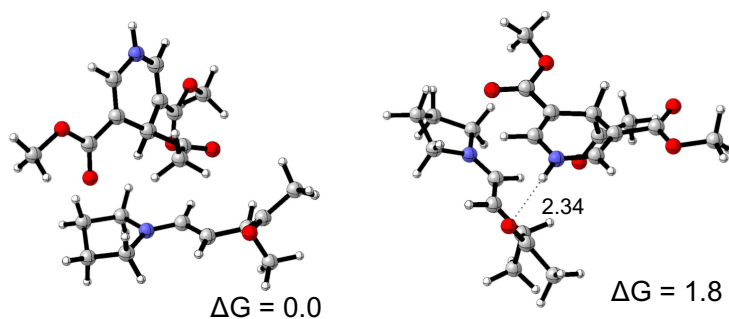


Figure 3.4. Low energy adduct-DHP dimer conformations

Based on the above results, the preferred mechanism is confirmed to be mechanism 1 as presented in Figure 3.2.A. Hence, the enantioselectivity-determining step is the radical addition between the iminium ion and the acyl (COR^2) radical. The rest of this chapter focuses on studying the enantioselectivity-determining step in detail.

3.4 Automations and conformation analyses

Traditionally, to obtain TS structures of complex systems, computational chemists need to follow a process of conformational searches, DFT optimisations to minima on the potential energy surface, bond freezing optimisations and TS optimisations.¹³⁶ In the case of catalyst C, the conformational searching calculations of its iminium produce 2228 conformers. Considering the size of the system (for example, 86 atoms and 388 electrons in the catalyst C iminium), it is computationally costly to re-optimize and obtain a TS for all the conformers at the DFT level from the output of conformational searches. Hence, an efficient procedure was established for studying the radical addition process computationally (Figure 3.5).

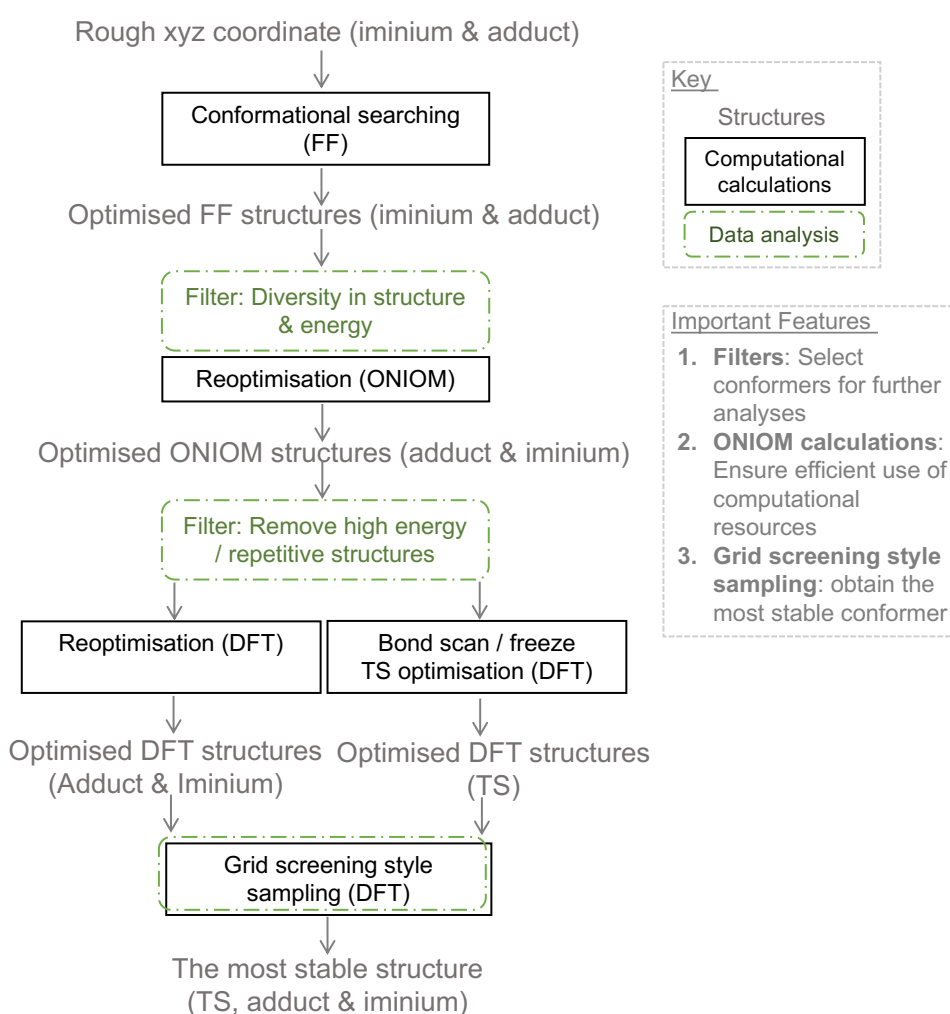


Figure 3.5. An efficient computational procedure for uncovering large and flexible chemical systems. The grid screening style sampling requires both computational calculations and data analyses.

Firstly, filters were introduced into the procedure, which aimed to select conformers for further analyses. Filters were built on top of a conformational labelling system. The conformational labelling system was developed to classify and label conformers according to their structural features. The procedure starts with the identification of the key structural features within the molecule. Figure 3.6 summarises the

derivations of structural feature labels. At the beginning of our investigation, only the iminium and rotamer features were included in the conformation label (Figure 3.6.A and 3.6.B). These two structural features are known to be correlated with the energy of the system.^{122,137} After the first batch of calculations, it was determined that a conformation label with two features is insufficient and conformers with the same conformation label can have a $\Delta\Delta G$ up to 6 kcal mol⁻¹. By inspecting the difference in structure between conformers with the same conformation label, other important structural features were picked out and introduced into the conformation labelling system. With the current system, $\Delta\Delta G$ values of conformers with the same conformation label should be under 1.0 kcal mol⁻¹.

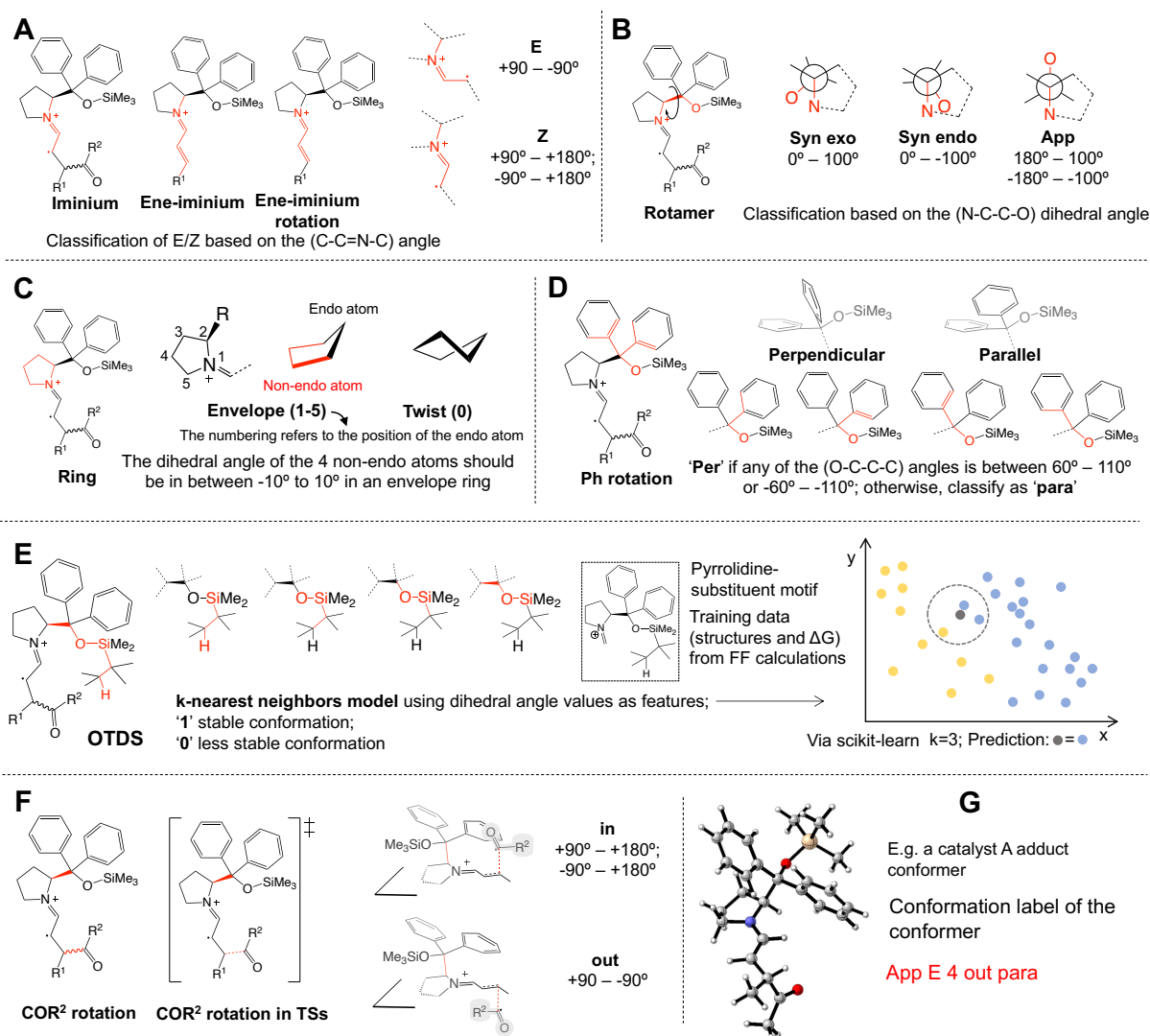


Figure 3.6. The conformational labelling system: The structural feature labels are identified based on dihedral angle values. The key dihedral angles are coloured in red. A) iminium, ene-iminium and its rotation; B) ring; C) rotamer (*ie* the rotation of the large bulky substituent); D) Ph rotation; E) -OTDS based on k-nearest neighbours model; and F) COR² rotation; G) The conformation label is derived by joining the structural feature labels identified.

The key structural features and the associated labels for the cationic radical adduct are as follows: iminium (E/Z; Figure 3.6.A), rotamer (syn_exo/syn_endo/app; Figure 3.6.B), ring conformation (envelope(1-5)/ring(0); Figure 3.6.C), Ph rotation (para/per; Figure 3.6.D) and COR² rotation (in/out; Figure 3.6.F). The rotamer, ring conformation and Ph rotation classifications are also applied to iminium structures. The differences between the adduct and the iminium are taken into account by the ene-iminium ion (EE/EZ/ZE/ZZ) and its rotation (trans-/cis-ene-iminium) structural feature (Figure 3.6.A). All the structural features for the iminium and the COR² rotation label are also applicable to the TS structures. The conformation label is derived by joining the structural feature labels identified (Figure 3.6.G).

The label identification process is based on measurements of dihedral angles. For example, the iminium (E/Z) structural feature classification is based on the value of the C-C=N-C dihedral angle (Figure 3.6.A). A conformer with an 'E' iminium has a $\theta(\text{C-C=N-C})$ value between $+90^\circ$ and -90° . Conformers with a $\theta(\text{C-C=N-C})$ value outside the range of $+90^\circ - -90^\circ$ are regarded as 'Z' iminiums. From dihedral angle distribution studies (Figure 3.7), none of the conformers has a dihedral angle similar to the cut-off. The label identification for other features follows a similar logic of categorical classification based on one or multiple dihedral angle values (see Appendix to Chapter 3, 3.B for more details).

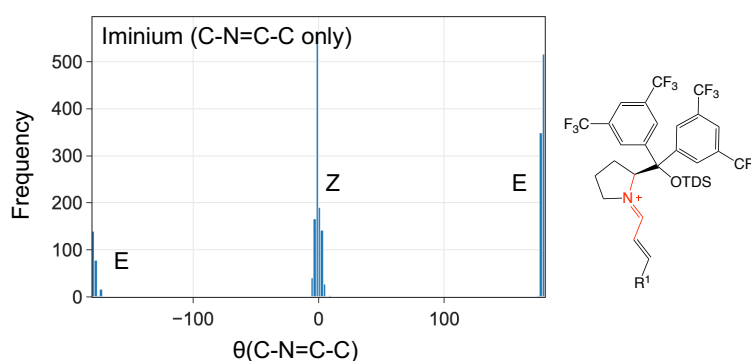


Figure 3.7. Dihedral angle distribution studies for the iminium (E/Z) structural feature with MMFF structures.

The derivation of the ring conformation feature is slightly more complex. The heterocyclic 5-membered ring can either adopt an envelope or a twist conformation (Figure 3.6.C). When only the combinations of adjacent atoms are considered, there are five distinct dihedral angles within a 5-membered ring. The values of the five dihedral angles were measured. For an envelope conformation, the dihedral angle of the four non-endo atoms should be close to 0° . Hence, a range between -10° to 10° was chosen as the criteria for identifying the endo atoms based on the dihedral angle value of the non-endo atoms. A label between '1' and '5' was given for conformers with an envelope ring, where the number refers to the endo atom position. If none of the five dihedral angles has a value between -10° and 10° , the conformer shall be classified as having a twist ring and given a label of '0'. The ring conformation label was not included in the conformation label of the conformer in filtering for selecting MMFF conformers to optimise at the ONIOM or DFT level. The ring has a very flat potential surface and the geometry

commonly changed after DFT optimisations.¹³⁸ The changes in conformation did not follow any pattern and were unpredictable.

Structures with catalyst A_OTDS and C have an -OTDS chain. With the dihedral angle values on the backbone of the -OTDS chain as features, a k-nearest neighbours model was trained based on force field (FF) data to label the stable -OTDS orientations (Figure 3.6.E). The dataset for training the model consists of 221 conformers of the pyrrolidine-substituent motif structure presented in Figure 3.6.E. In labelling the conformers from the training dataset, conformers were separated according to their 'rotamer' (syn_exo/syn_endo/app) label. Within each data group, the ΔG value of the conformers was normalised to a value between 0 and 1. The three sets of data were then combined. A label of '1' (*ie* a stable structure) was given if the normalised ΔG of the conformer was less than 0.2. Otherwise, a label of '0' (*ie* a less stable structure) was given. In additionally, all conformers were re-optimised at the ω B97X-D/6-311G(d,p)//B3LYP/6-31G(d) level of theory, and the k-nearest neighbour model training was repeated using this dataset. The model based on data from DFT calculations produces results similar to those from FF calculations.

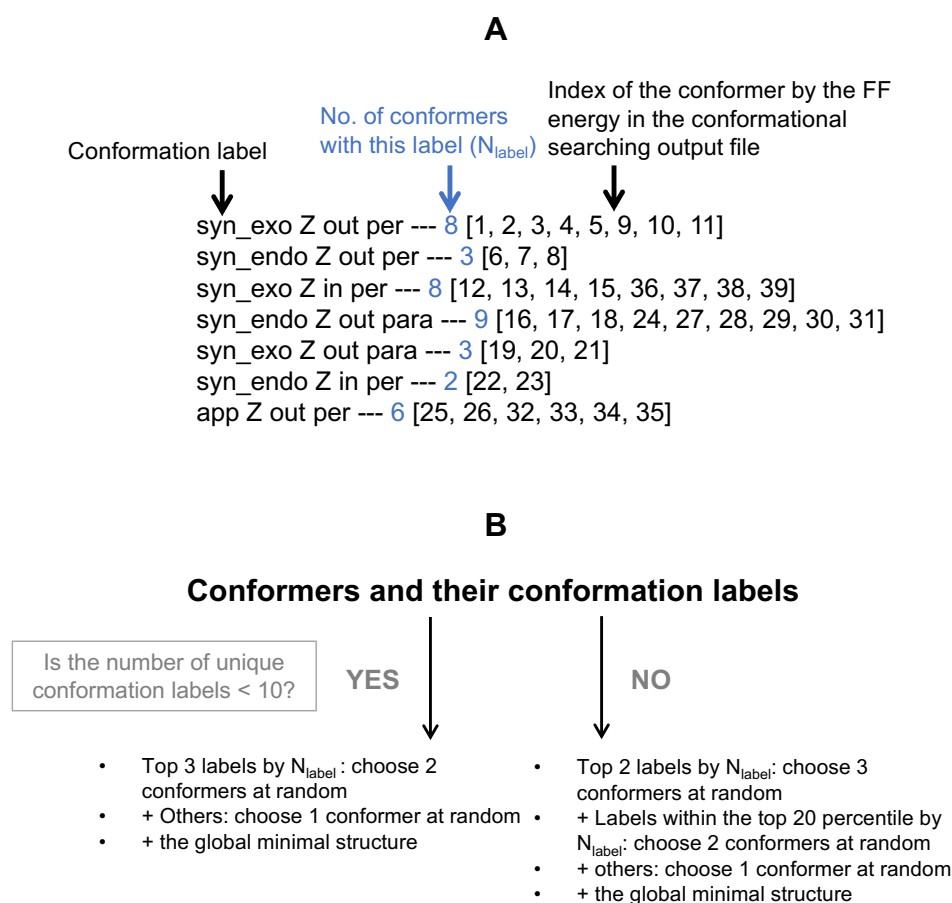


Figure 3.8. Using conformation labels to select conformers from conformational searching output files. A) Outputs from the conformation label generation; B) The workflow for picking conformers for further analyses based on their conformation labels

The automated filter pipeline (Figure 3.5) generates the conformation label with the xyz file of the conformers and dihedral angles of interest as the input. Based on the conformation label and the number of conformers under the same classification, conformers are selected for further analyses, *ie* optimisations at a higher level of theory or TS optimisations (Figure 3.8). The filter on conformers from conformational searches aims to select representative structures that are diverse in structure and energy for re-optimisations at a higher level of theory with ONIOM calculations. Taking catalyst C iminium as an example, 63 representatives are selected from 2228 MMFF structures for ONIOM optimisations.

ONIOM optimisations¹³⁹ were performed on structures from conformational searches at the ONIOM(B3LYP/6-31G(d):UFF) level of theory, *ie* part of the system is optimised at the cheaper UFF (universal force field) level.¹⁴⁰ ΔG values from ONIOM calculations and at the DFT level correlate well if the border of the two layers were carefully chosen (Figure 3.9). On average, based on the calculations on SR adducts with catalyst C and B, the computational time for ONIOM calculations is only 1/3 of DFT level calculations typically. Thus, initial sampling with ONIOM calculations allows a large number of conformers to be screened and studied at a relatively low computational cost. The lowest energy conformers in the group of conformers with the same conformation label and all structures with a $\Delta G < 5$ kcal mol⁻¹ were selected for DFT re-optimisations.

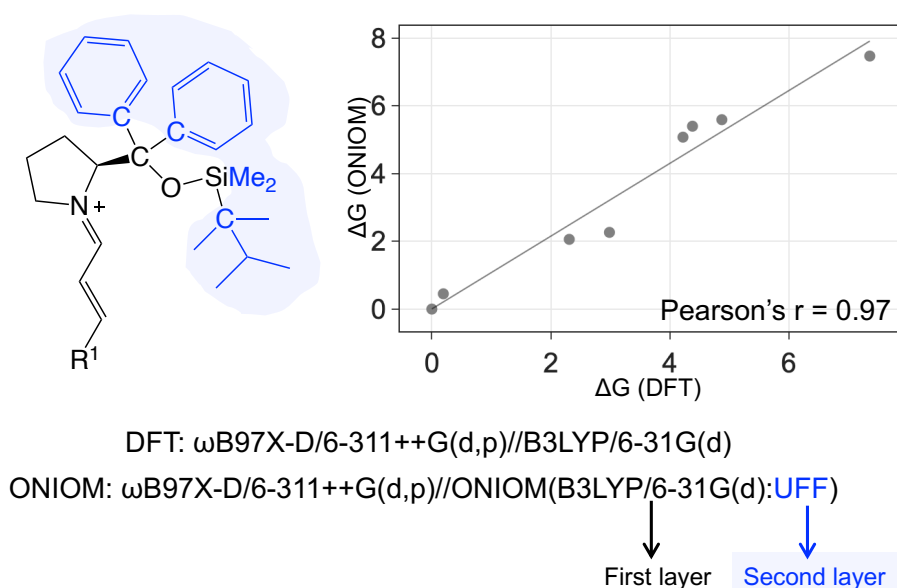


Figure 3.9. The border division for ONIOM calculations. In the structural diagram, the core DFT layer is black and the UFF layer is blue. The ONIOM level ΔG values of the conformers are plotted against their DFT level ΔG values. (Unit: kcal mol⁻¹)

Finally, grid screening style samplings were conducted to ensure that the lowest energy structure was considered and obtained. Here, the data from the conformation labelling system was used as a guide for further conformation sampling. The process began with exploring correlations between structural features and ΔG of the system with statistical analyses through box plots (Figure 3.10.A). The box plot

analyses allowed generalisations to be made. For example, conformers of the catalyst B TS with the 'syn_exo' and 'per' labels are noticeably more stable than conformers with 'app', 'syn_endo' and 'para' labels. Sometimes, more than one structural feature label may contribute to a stable conformation. In the case of catalyst B TSs, '3' vs '4' (ring conformation), 'EE' vs 'EZ' (ene-iminium) and 'in' vs 'out' (COR rotation) are competing structural features in contributing to a low energy conformation. Additional structures with different combinations of these labels were sampled in a grid screening style if they were not already presented in the dataset of optimised structures (Figure 3.10.B).

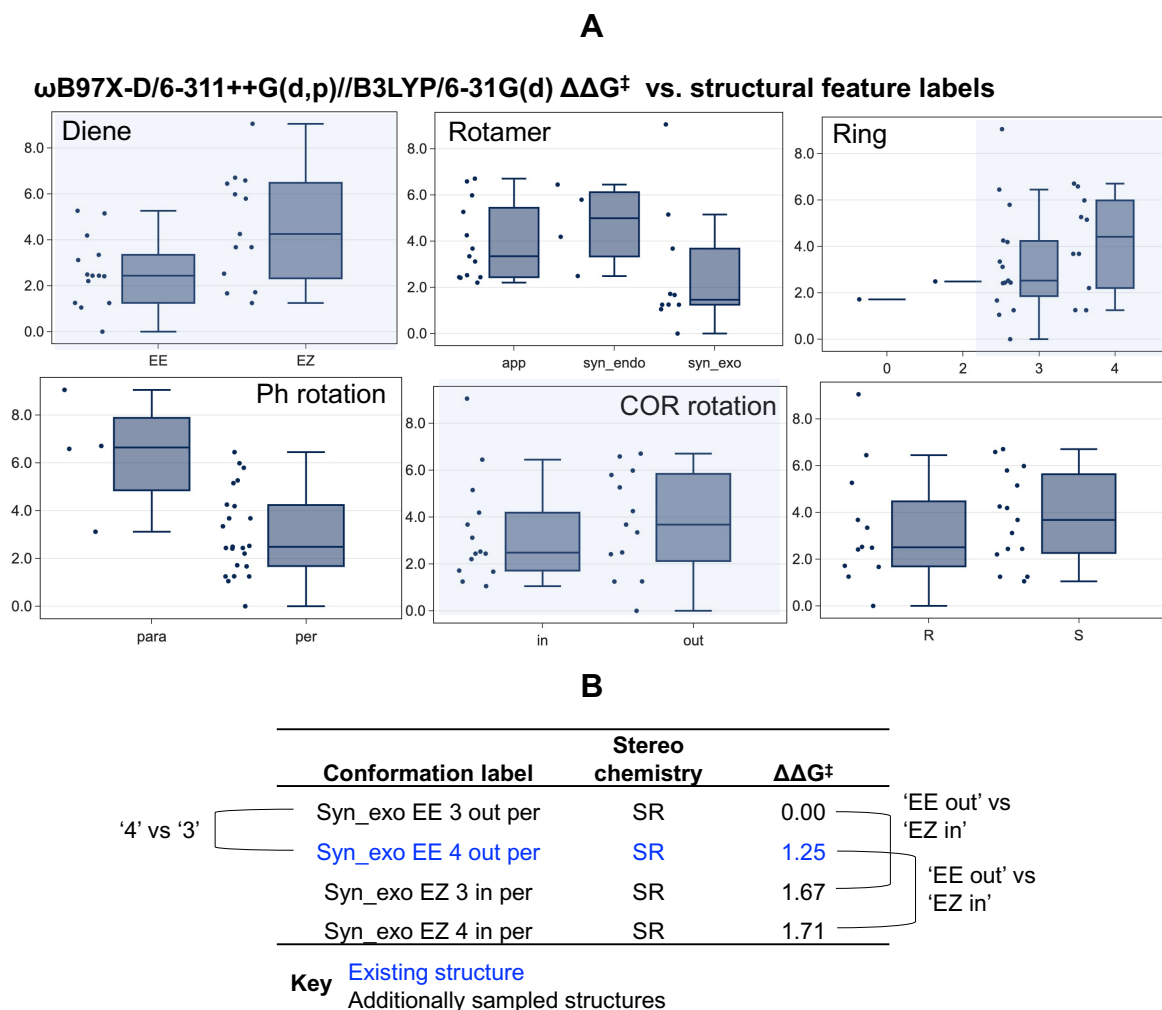


Figure 3.10. Grid screening style samplings. A) Box plots of $\Delta\Delta G^\ddagger$ values vs key structural feature labels TSs with catalyst B for the radical addition step, *ie* the energy-conformation profile: The actual data points are included next to the box plots in each graph. The energetically ‘competing’ structural features have been highlighted in blue. B) Further samplings at the DFT level: Structures in blue already existed in the dataset before the grid screening style sampling. The structures in black were the newly sampled data points and the input to the calculations was derived from existing optimised structures.

The improved procedure allows efficient use of computational resources and is effective in obtaining the key structures for analyses. This procedure was developed during the investigation on reactions with catalyst A and A_OTDS and has been successfully applied on the more complex systems with catalyst B and C.

3.5 Origin of the enantioselectivity

With the automated procedure described in the last section (Figure 3.5), various systems have been studied thoroughly by considering catalysts (*ie* Catalyst A, A_OTDS, B and C; Figure 3.1) and substrates (*ie* $R^1 = \text{Ph}$ or Me and $R^2 = \text{Ph}$ or Me) of different size and complexity. This section of the chapter focuses on the enantioselectivity-determining step, the radical addition between the iminium ion and the acyl (COR^2) radical (Figure 3.11).

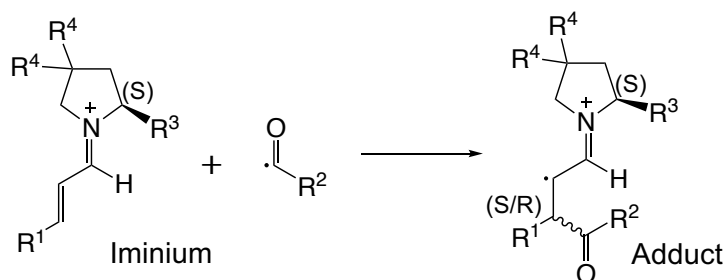


Figure 3.11. The enantioselectivity-determining step of the reaction, the focus of this section. Experimentally, the *R* product was obtained as the major product. In this study, $R^1 = \text{Ph}$ or Me and $R^2 = \text{Ph}$ or Me . R^4 and R^3 depend on the specific catalyst.

The computational $\Delta G^\ddagger$ and the conformation label of the lowest energy conformations for selective systems are presented in Table 3.1 and 3.2 (see Table S3.1 and S3.2 for the complete result). The trend in $\Delta\Delta G^\ddagger$ generally agrees with the experimental results. In addition to the data from Melchiorre *et al.*, the results were also compared with the work from Yu *et al.*¹¹⁸ Yu *et al.* have also reported a highly similar radical route for synthesising 1,4-dicarbonyls. Their reactions also involved amine catalysts but used Ru(I) complex for photoinitiations and generations of COR^2 radicals. Pathways via a *SR* TS lead to the major product. In Table 1, as the steric demand of the catalyst increases, *SR* pathways become more kinetically favourable than *SS* pathways. Changing R^1 from $-\text{Me}$ to $-\text{Ph}$ group also contributes to a higher $\Delta\Delta G^\ddagger$. It is important to highlight that $\Delta G^\ddagger$ values of *SR* pathways remain almost unchanged regardless of the changes in the steric bulkiness and electronic properties of the catalyst. This suggests that the radical addition step is not the yield-determining step.

Table 3.1. The calculated kinetic data (level of theory: ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d); unit: kcal mol⁻¹) and experimental percentage enantiomeric excess (%ee) from the work of Melchiorre *et al.*⁴ and Yu *et al.*¹¹⁸

	R¹ = Me; R² = Ph			R¹ = Ph; R² = Ph		
	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	Exp. %ee	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	Exp. %ee
Catalyst A						
SR	12.2	0.5	-	-	-	12 ^b
SS	11.7	0.0	-	-	-	-
Catalyst A_OTDS						
SR	11.6	0.0	-	-	-	-
SS	12.0	0.4	-	-	-	-
Catalyst B						
SR	11.8	0.3	-	12.4	0.0	27 ^a /38 ^b
SS	11.5	0.0	-	14.7	2.3	-
Catalyst C						
SR	11.6	0.0	69 ^a	11.2	0.0	76 ^a /80 ^b
SS	13.6	2.0	-	13.7	2.5	-

a. From Melchiorre *et al.*⁴

b. From Yu *et al.*¹¹⁸

Table 3.2. The conformation label of the iminium in the lowest energy TS conformer and the conformation label of the lowest energy ground state iminium conformer at the ω B97X-D/6-311++G(d,p)/SMD /CH₃CN//B3LYP/6-31G(d) level of theory. All the structures presented below have a 'trans-ene-iminium' label.

Catalyst	SR TS	SS TS	Iminium
R¹ = Me; R² = Ph			
A	syn_exo EE 3 per	syn_exo EE 3 per	syn_exo EE 3 per
A_OTDS	syn_exo EE 3 per 1	syn_exo EE 3 per 0	syn_exo EE 3 per 1
B	syn_exo EE 3 per	syn_exo EZ 3 per	syn_exo EE 3 per
C	syn_exo EE 4 per 1	syn_exo EE 4 per 1	syn_exo E 0 per 1
R¹ = Ph; R² = Ph			
B	syn_exo EE 3 per	syn_exo EZ 3 per	syn_exo EE 3 per
C	syn_exo EE 0 per 1	syn_exo EZ 5 per 1	syn_exo EE 0 per 1

A benchmarking study has been conducted on the key TS structures. Firstly, the results highlight the importance of including a solvent model in single-point calculations, which improves the accuracy of the TS $\Delta G^\ddagger$. $\Delta\Delta G^\ddagger$ values show a closer match to the experimental result. See Appendix Table S3.3 and S3.4 for results from calculations without the solvent model. Secondly, there are often several TS conformers with very similar $\Delta G^\ddagger$ values to the lowest energy TS, especially for the SS TSs. Hence, the lowest energy SS TS conformer obtained with different DFT functionals may have subtle structural differences. Nevertheless, the overall trend stated in the previous paragraph remains consistent.

Intrinsic reaction coordinate (IRC) or quick reaction coordinate (QRC)¹⁴¹ calculations are performed to check that the TS is on the reaction path of interest. Further analyses were conducted taking into consideration of all the optimised TS conformers by plotting $\Delta\Delta G^\ddagger$ of the TSs against ΔG of the iminium

taken from the IRC or QRC of the corresponding TS (Figure 3.12). For catalyst A, a strong linear trend (Pearson's $r = 0.94$) in the plot between the two variables shows that the stability of the iminium ion is strongly correlated to the activation energy of the process. Increasing the complexity and steric demand of the catalyst leads to a drop in the Pearson's r value of the plot. The gradient and y-intercept of the line of best fit also deviate from 1.0 and 0.0. The trend implies that the iminium in TSs with a low $\Delta\Delta G^\ddagger$ may take up a less stable conformation as the steric hindrance increases.

Looking at the lowest energy TS structures, the iminium ion in SR TSs always has the same conformation label (EE) as the lowest energy ground state iminium (Table 3.2). In the radical addition process for the SR adducts, the COR² radical can simply attack with the 'out' orientation without interfering with the bulky substituent in the iminium. However, this does not apply to the SS TSs. The iminium ions in the lowest energy SS TSs often need to adopt a less stable EZ conformation, unlike the ground state iminium. The difference in iminium conformation may be due to local interactions or structural deformations during the reaction process.

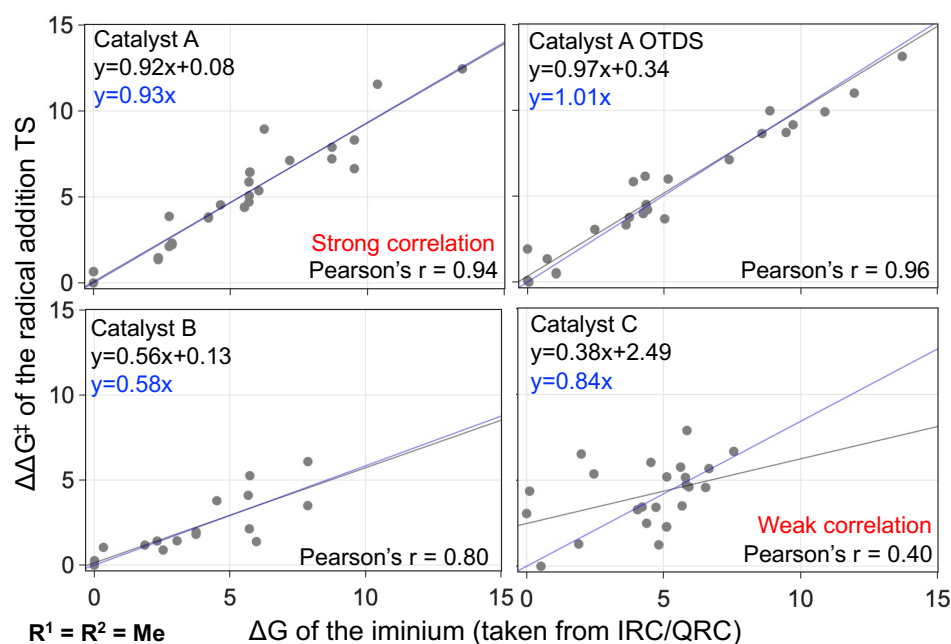


Figure 3.12. $\Delta\Delta G^\ddagger$ values of the TSs vs ΔG of the iminium taken from the IRC or QRC¹⁴¹ plots for the radical addition process in reactions with catalyst A, A_OTDS, B and C. The substituents on the enal and COR² radical, R^1 and R^2 , are both Me group (Figure 3.11). The calculations were performed at the ω B97X-D/6-311G(d,p)//B3LYP/6-31G(d) level of theory. (Unit: kcal mol⁻¹)

Local interactions refer to either favourable non-covalent interactions (NCIs) or unfavourable steric clashes. NCI plot analyses provide visualisations of the position and strength of NCIs within a structure and were conducted on the TSs.¹⁴² Various TS conformers, including the less stable conformers, for processes with different catalysts have been studied and consistent results have been obtained. The NCIs between the COR² radical and the iminium ion are mainly weak van der Waals interactions. In Figure 3.13, examples of NCI plot structures are given. For NCI plots of other systems, please see Appendix Figure S3.4. The van der Waals interactions are indicated by green surfaces within the structures. Local interactions do not play a significant role in determining the orientation of the iminium in the lowest energy SS TS, regardless of the size of the substituent on the catalyst.

Catalyst C; R¹ = R² = Ph systems

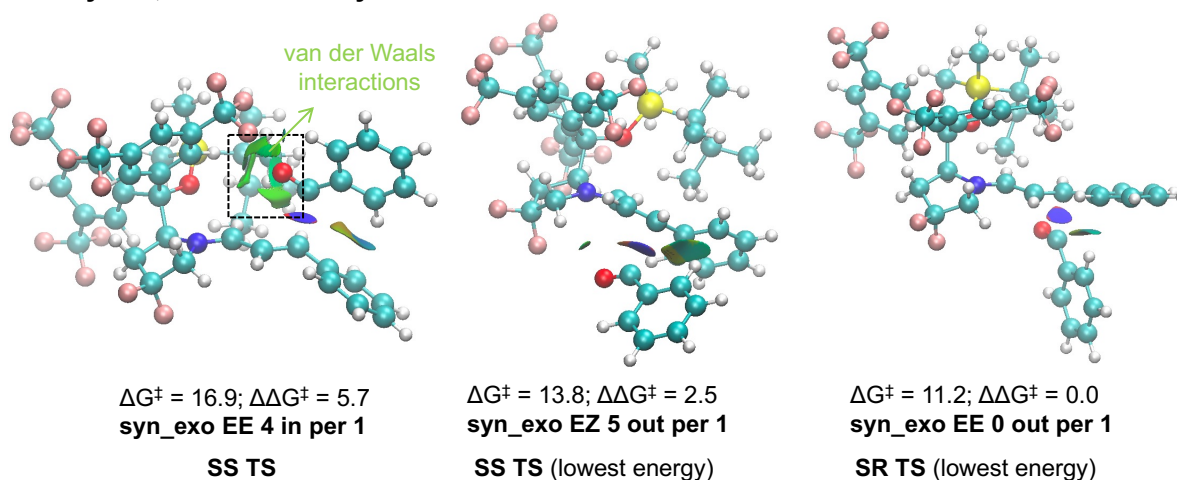


Figure 3.13. Non-covalent interactions (NCIs) analyses on the TSs. The colour of the surfaces within the TS structures indicates the nature of the interactions (*ie* blue = attractive, red = repulsive and green = weak van der Waals interactions). The results presented above are for catalyst C (R¹ = R² = Ph) TS systems. The lowest energy TSs with the specific stereochemistry are labelled in the diagram. (Level of theory: ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d), unit: kcal mol⁻¹)

The overall structural deformation, as opposed to specific rotational conformational changes, is considered with distortion/interaction analyses (Figure 3.14.C).¹⁴³ $\Delta U_{d,iminium}^\ddagger$ is the difference in U between the iminium structure from TS and the ground state iminium structure and measures the extent of distortion between the iminium structure from the TS relative to the ground state iminium from QRC in terms of energy. $\Delta U_{d,iminium}^\ddagger$ values were calculated at the ω B97X-D/6-311++G(d,p) level of theory and compared between systems of different complexity (Figure 3.14.A and B and Appendix Table S3.5). In catalyst C systems, the $\Delta U_{d,iminium}^\ddagger$ is much higher when the iminium in the SS TS takes up the same conformation label, EE, as the lowest energy ground state iminium. The difference in the extent of deformation is also reflected in the alignment diagrams in Figure 3.14.B. The iminium in the preferred SS TS takes up a less stable EZ conformation to minimise the overall deformational distortion. The change in the iminium conformation is at the cost of raising the kinetic barrier of the radical addition process, which leads to the stereo outcome of the reaction.

For simple systems with a less sterically bulk catalyst (for example, catalyst A system; $R^1 = R^2 = \text{Me}$; low experimental %ee; Figure 3.14.A), the lowest energy SS and SR TS have the same conformation label. Here, the energy cost of distorting the structure is presumably lower than the cost of changing the iminium conformation. For catalyst A ($R^1 = \text{Me}$; $R^2 = \text{Me}$), the corresponding $\Delta U_{d,iminium}^\ddagger$ for the lowest energy SS and SR TS differs by only 1.0 kcal mol⁻¹. Additionally, $\Delta U_{d,iminium}^\ddagger$ values also correlate positively with $\Delta\Delta G^\ddagger$ of the corresponding TS (See plots of $\Delta U_{d,iminium}^\ddagger$ vs $\Delta\Delta G^\ddagger$ in Figure 3.15).

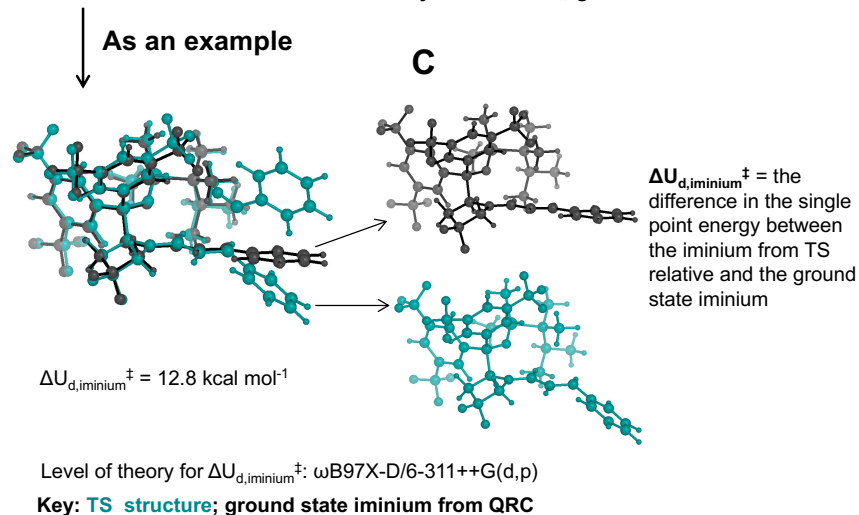
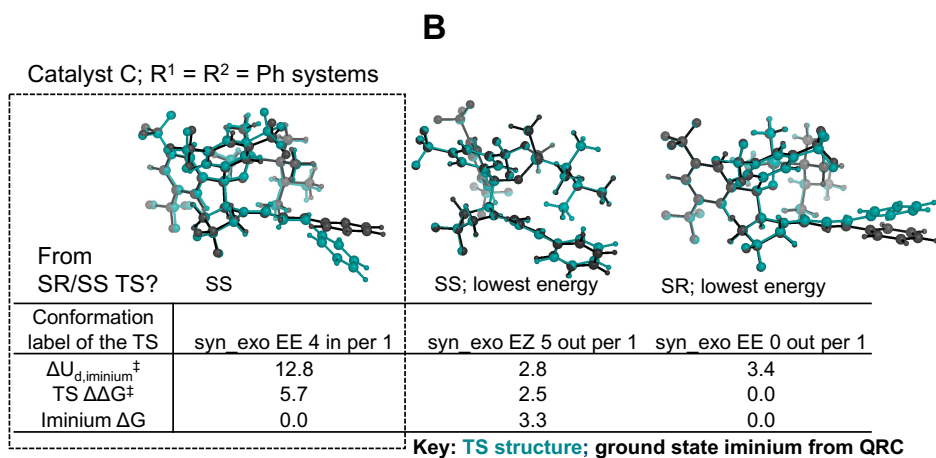
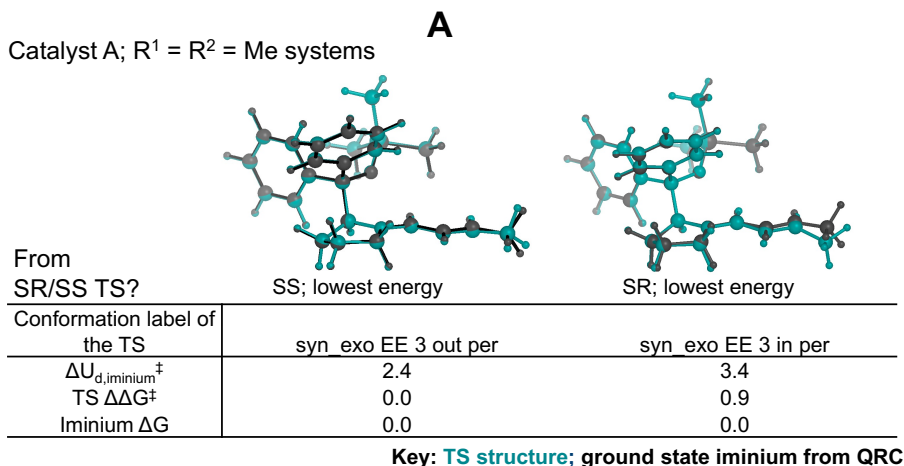


Figure 3.14. Distortion-interaction analyses for the iminium structure from TSs of the radical addition process. Structural alignment diagrams between the iminium structure at the TS and the ground state iminium are also given. The conformation label of the TSs is included in the table. All the TSs presented have a 'trans-ene-iminium' label. The lowest energy TSs with the specific stereochemistry are labelled in the diagram. TS $\Delta\Delta G^\ddagger$ and iminium ΔG values are reported at the $\omega B97X-D/6-311++G(d,p)/SMD/CH_3CN//B3LYP/6-31G(d)$ level of theory. $\Delta U_{d,iminium}^\ddagger$ values are reported at the $\omega B97X-D/6-311++G(d,p)$ level of theory. (Unit: kcal mol⁻¹)

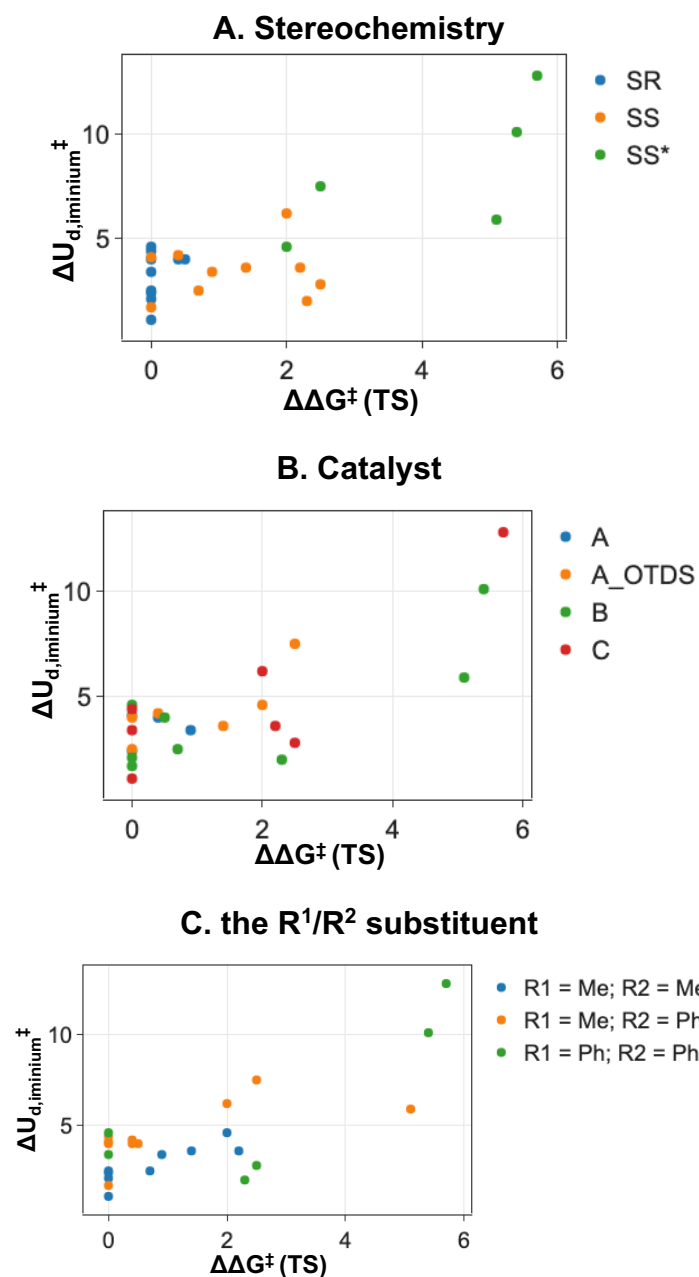


Figure 3.15. $\Delta U_{d,iminium}^{\ddagger}$ vs change in activation energy ($\Delta\Delta G^{\ddagger}$) plots. $\Delta U_{d,iminium}^{\ddagger}$ measures the extent of distortion between the iminium structure from the TS relative to the ground state iminium from QRC in terms of energy. The data points are coloured by A) stereochemistry (*ie* SS or SR); B) the catalyst involved (*ie* A, A_OTDS, B and C); C) the R¹/R² substituent. 'SS*' refers to SS TSs where the iminium ion in the TS adopts the same conformation as the lowest energy ground state iminium. TS $\Delta\Delta G^{\ddagger}$ values are reported at the ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d) level of theory. $\Delta U_{d,iminium}^{\ddagger}$ values are reported at the ω B97X-D/6-311++G(d,p) level of theory. (Unit: kcal mol⁻¹)

Changing from -OTMS to -OTDS always led to a noticeable increase in %ee and $\Delta\Delta G^\ddagger$, regardless of the complexity of the system (*ie* from catalyst A to A_OTDS or from catalyst B to C systems) (Table 3.1). Introducing the -CF₃ groups only (*ie* moving from catalyst A to B systems) gave similar %ee and $\Delta\Delta G^\ddagger$ values. However, having both the -OTDS chain and -CF₃ groups in the catalyst (*ie* moving from catalyst A_OTDS to C systems) produced the best enantioselectivity outcome. The lowest energy TSs take up a syn_exo orientation, where the -CF₃ substituted phenyl groups are on average further away from the site of radical addition compared to the -OTDS chain. Thus, -CF₃ groups indirectly influence the reaction site by reducing the steric flexibility of the -OTDS chain. The enhanced structural restrictions force iminium ions to take up a less stable structure at the TSs. The electronic impact of introducing the -CF₃ groups is insignificant. Hirshfeld charge analyses show that the charge distribution at the site of radical additions (*ie* the N=C-C=C region) does not vary extensively between iminium ions with different catalysts (Figure 3.16).

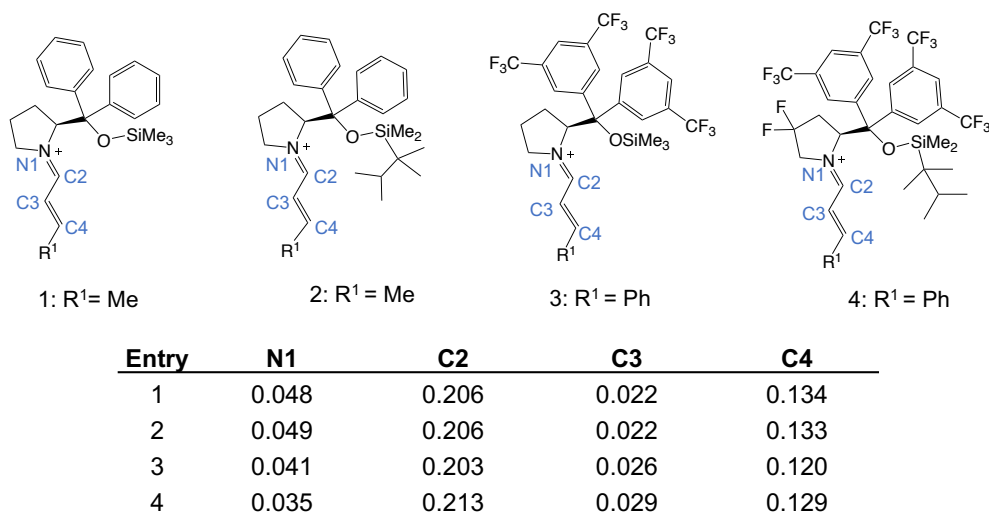


Figure 3.16. Hirshfeld Charge analyses on the iminium ion. The analyses are performed at the ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d) level of theory.

In addition to the work from Melchiorre *et al.*⁴ and Yu *et al.*,¹¹⁸ the experimental results from other amine-catalysed reactions via a radical pathway have also been considered.^{124–128,144} In general, increasing the steric complexity of the catalyst improves the selectivity of the reactions. The above model and lesson learnt can also be applied to similar reactions given that the substrates (*ie* attacking radical and enal, ketone or aldehyde) are generally inert to form strong non-covalent interactions.

3.6 Summary

In conclusion, an effective procedure for computational investigations on the reaction of interest has been proposed and assembled in this study. The ideas behind the procedure can be transferrable and applied to other large, structurally complex and flexible organic systems.

The photochemical synthesis of 1,4-dicarbonyl compounds with amine catalysts has been studied computationally with DFT and FF calculations. Focusing on the radical addition step between the iminium and the COR² radical, the origin of enantioselectivity becomes clear. The direction of attack of the COR² radical and its interactions with the bulky substituent on the catalyst are not sufficient to explain the difference in $\Delta G^\ddagger$ value. The conformation labelling system and distortion/interaction analyses highlight that the iminium in the preferred SS TS adopts a less stable EZ conformation to reduce the overall structural deformation. The enantioselective outcome of the reaction is due to the conformation difference between the iminium in the SR and SS TS, which contributes to the $\Delta\Delta G^\ddagger$ value significantly. For systems with reduced substituent size in the catalyst, the iminium adopts the same structure in SS and SR TSs (Figure 3.17). The small structural deformation difference between the iminium in the TSs leads to insignificant $\Delta\Delta G^\ddagger$ values and poor enantioselectivity experimentally.

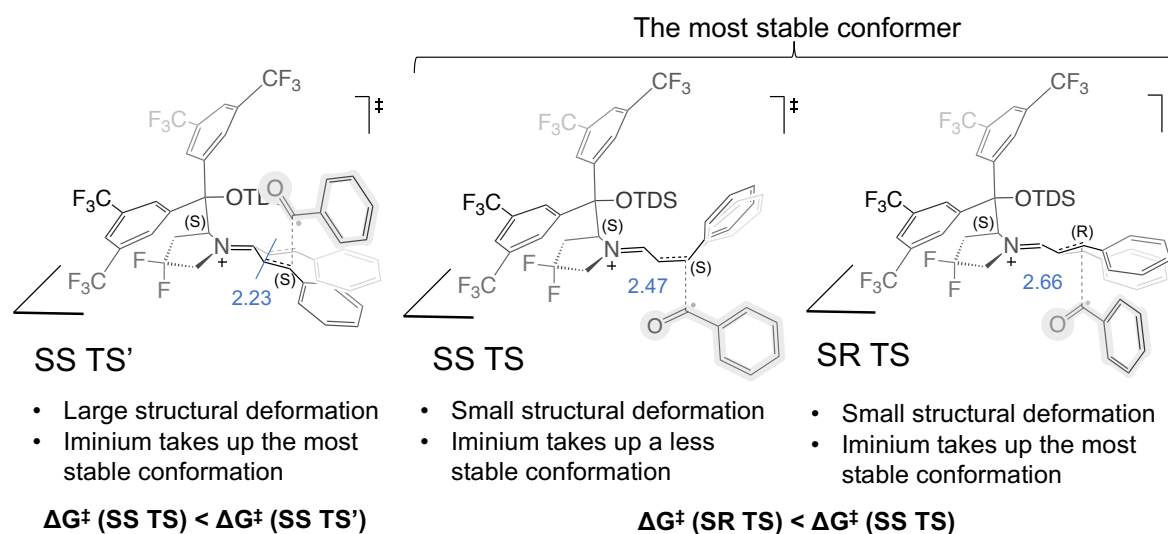


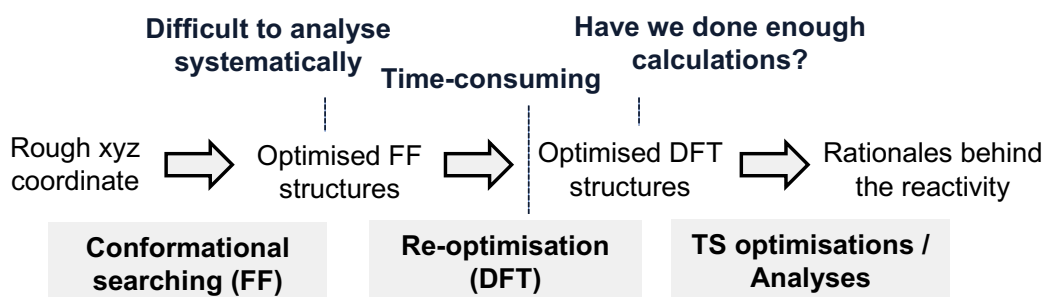
Figure 3.17. The origin of enantioselectivity: Catalyst C TSs ($R^1 = \text{Ph}$; $R^2 = \text{Ph}$) are used as examples. The iminium in the SR TS takes up the lowest energy conformation with little structural deformations. In SS TS, small structural deformations within the iminium are at the cost of changing the conformation. Thus, $\Delta G^\ddagger$ (SR TS) < $\Delta G^\ddagger$ (SS TS). In SS TS', the iminium takes up the lowest energy conformation, which leads to large structural deformations. Hence, $\Delta G^\ddagger$ (SS TS) < $\Delta G^\ddagger$ (SS TS'). The bond length values of the forming C-C bonds in Å are labelled in blue. The SS and SR TS lead to the SS and SR adduct respectively.

Chapter 4

CONFPASS: fast DFT re-optimisations of structures from conformation searches

4.1 Background

Making the DFT re-optimisation process of structures from conformational sources more efficient is a challenge. In Chapter 3, an automated pipeline has been built to select representative conformers for re-optimisations using a conformation labelling system based on dihedral angle values. The procedure was specifically designed for studying the reactivity of diarylprolinol silyl ether catalytic systems. This chapter is on the development of CONFPASS (**C**onformer **P**rioritisations & **A**nalysis for DFT re-optimisations), an open-source Python package available on GitHub (github.com/Goodman-lab/CONFPASS). CONFPASS builds on the approach in Chapter 3 to generate a robust pipeline applicable to diverse organic structures.



CONFPASS offers:

- Analysis on the conformational searching outputs and a priority list for re-optimisations at a higher level of theory
- Prediction on the completion of the re-optimisation process with confidence

Figure 4.1. The bottleneck in the computational workflow for understanding chemical reactivity. CONFPASS provides a solution.

DFT optimisations of the FF level structures from conformational searches are a bottleneck in the computational workflow for understanding chemical reactivity (Figure 4.1). How can we analyse conformer structures from conformational searching calculations systematically? To take a simple example: a molecule with ten FF conformations. If the lower energy five are re-optimised using DFT, is it necessary to re-optimize the other conformations? CONFPASS provides an answer to these questions.

CONFPASS starts with the output from an FF-based conformational searching program, with the conformers sorted in order of increasing FF energy. If these are all re-optimised at DFT level to generate

a new list, ordered by DFT energy, the new list may well contain fewer structures, as different FF structures may optimise to the same DFT structure (Figure 1.4). Whilst structures with high FF energies may re-optimize to structure with energies closer to the DFT global minimum, structures with very high FF energies may be safely discarded.

Several strategies can be used for deciding on the order for DFT re-optimisations. A simple but time-consuming approach, the *ascent* approach, is to re-optimize the structures according to the energy at the FF level from the lowest to the highest. This can be made faster by just selecting every n^{th} conformer from the FF output file, the *every n^{th}* approach. When $n = 1$, this is the same as the *ascent* approach. Another approach is to randomly pick a sample of conformers for re-optimisations, the *random* approach. For large and flexible systems, the last two approaches risk missing out key parts of the conformational space. Several studies have proposed methods to balance speed with confidence in the outcome. For example, Johannessen *et al.* process conformational sampling outputs with principal component analysis and clustering algorithm to assist Raman optical activity spectral analyses of vancomycin, a glycopeptide.¹⁴⁵ Merz Jr. *et al.* have developed AutoGraph, which clusters conformers and produces conformational graphs for metabolites, such as oxidised nicotinamide adenine dinucleotide.¹⁴⁶

CONFPASS chooses the optimal order in which to re-optimize the FF structures based on a clustering approach and then assesses how much confidence can be had in incomplete calculations as the re-optimisation progress. Thorough evaluations based on a dataset of structurally diverse molecules have shown that CONFPASS is effective in assisting organic chemists explore conformational space and structures.

4.2 Data gathering

This study is based on a set of 822 molecules (Figure 4.2.A), 711 of which come from a dataset gathered by Hutchison *et al.*¹⁴⁷ An additional 91 molecules are hydrocarbons derived from the Hutchison molecules by substituting heteroatoms with carbon atoms in the SMILES string. Twenty molecules are taken from the work of Grayson *et al.*⁵² The Grayson dataset molecules are mostly organic catalysts with hydrogen-bonding donating characteristics. Additionally, a small number of molecules, including three radical structures, come directly from literature.^{148–151} The structures are listed in Appendix Figure S4.1. Conformational searching calculations were conducted on the 822 molecules in MacroModel (v11.7)³⁹ with Merck Molecular Force Field (MMFF)¹⁵. Conformers with energies more than 10 kcal mol⁻¹ above the global minimum were omitted. For radicals, ground state pseudo structures were used for conformational searching calculations. A conformational searching calculation was considered complete when the following criteria were met:

1. No new global minimum structure was found during the last 0.25*total number of steps (*ie* the last 500 steps in the case of a calculation with 2000 steps).
2. The unique structures over the total number of steps should be less than 0.7.

The above criteria can be checked via inspections of the log file. If the calculation were considered incomplete, conformational searching calculations would be repeated with the global minimum structure from the previous calculation and with an increasing number of steps (*ie* from 2000 up to 4000 steps).

Out of the initial pool of 822 molecules, a subset of 150 was selected for the construction of the DFT dataset (Figure 4.2.B and 4.3). The 150 molecules include the 20 Grayson dataset molecules and 3 radical molecules. The remaining 127 molecules were randomly selected. Out of the 150 molecules, 22 are charged, with 2 of them being zwitterions. For this subset, all the FF structures were re-optimised from the conformational searching calculations at the DFT level. DFT optimisations were performed with Gaussian 16 (Revision B.01).¹⁰⁴ Here, the B3LYP-D3/6-31G(d) level of theory was chosen for the optimisation and frequency calculations.^{105–108} Single-point energy calculations were conducted at the ω B97X-D/6-311G(d,p) level of theory.¹⁰⁹ Unless specified, the 'DFT level' in the text below refers to this level of theory. The 150 molecules are diverse in terms of molecular size and flexibility. The total number of conformers ranges from 2 to 321 and the average number of atoms is 44. About 15 % of the re-optimised structures were found to be duplicates.

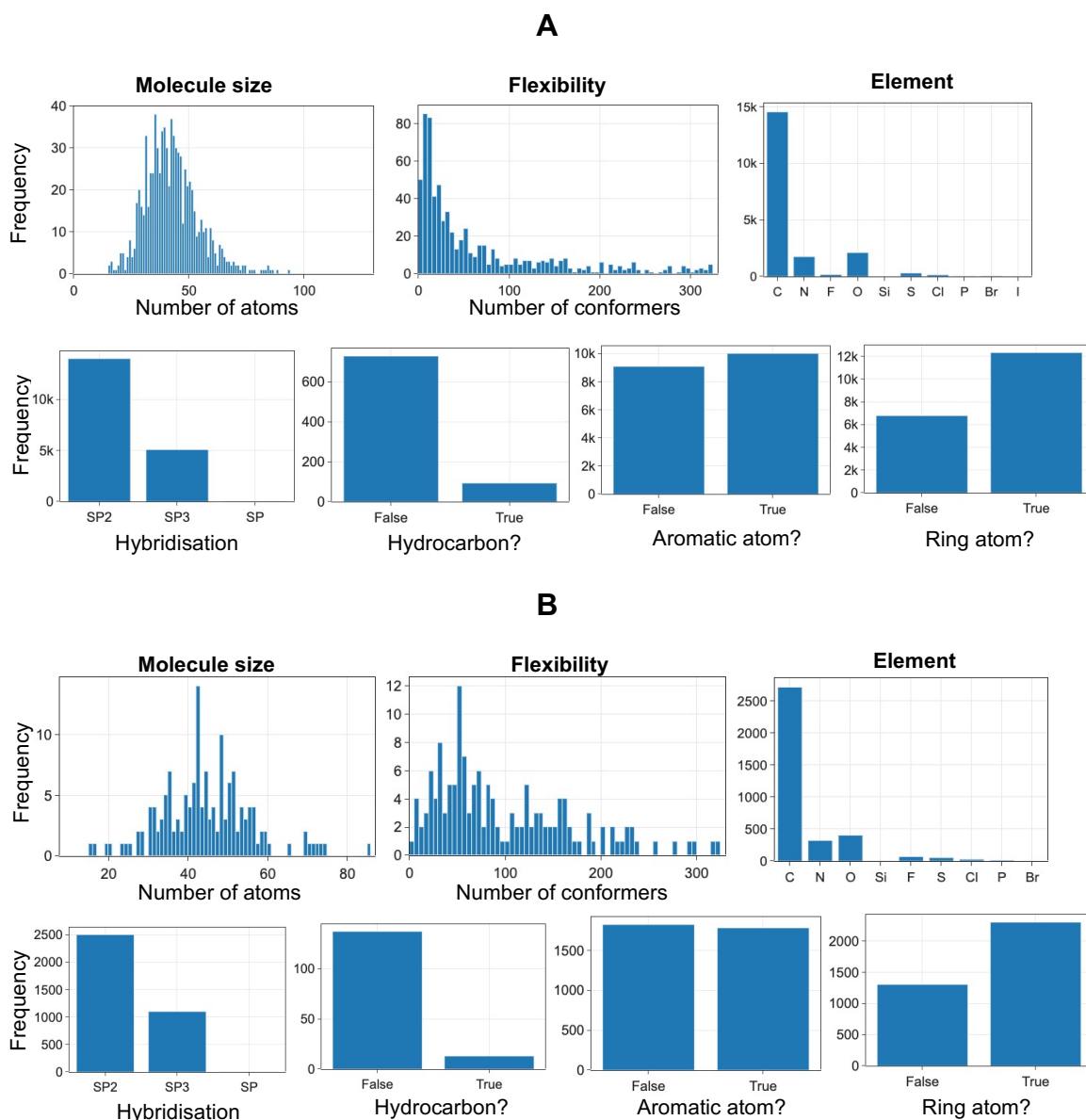


Figure 4.2. An overview of the A) 822 molecules in the conformational searching output dataset and B) 150 molecules in the DFT dataset. The histograms show that the organic molecules in both datasets are generally flexible and structurally diverse in terms of molecular size. The bar charts summarise the nature of the atoms in the molecules, which were determined by RDkit functions.

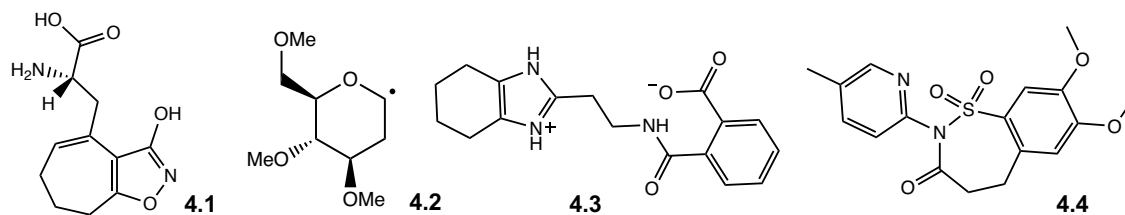


Figure 4.3. Example molecules in the DFT dataset.

4.3 Generating CONFPASS priority lists

Dihedral angle data for each conformation are used as descriptors for the clustering process. The CONFPASS priority list are generated based on results from the clustering process.

4.3.1 Dihedral angles descriptors

A list of bonds and bond order matrixes with and without H atoms are extracted from each structure. Bonds under the following categories are eliminated:

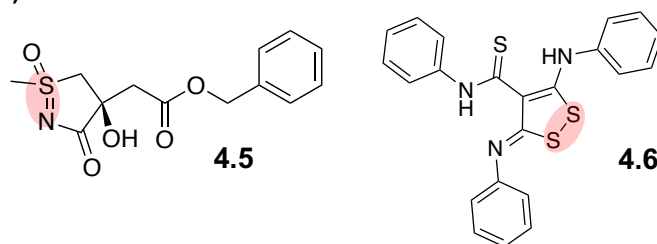
- A.** Bonds to monovalent atoms (F, Cl, Br, I, H);
- B.** Bonds to CX₃ groups (X = F, Cl, Br, I, H);
- C.** Bonds in small rings, *ie* 3-membered ring bonds
- D.** Bonds with bond order > 1, except for:
 - 1. C-X bonds with bond order = 1.5; (X = heteroatoms)
 - 2. X-X bonds with bond order = 2.0; (X = heteroatoms)
 - 3. C=N bonds (not within a ring)
 - 4. C-C(-Y) bonds with bond order = 1.5 (Y = O, S or N)

For these bonds, the dihedral angle (θ) values across all the conformers should have a small standard deviation (σ) and range (η) value. The σ and η parameters were calculated for 9017 sets of dihedrals with a bond order > 1 in the centre bond. The 9017 sets of dihedrals were identified using RDKit from 711 molecules in the Hutchison dataset.¹⁴⁷ Box plot analyses were performed on the statistical parameter and the outliers were inspected (Figure 4.4 and 4.5). Certain dihedrals with bonds containing heteroatoms or neighbouring to heteroatoms with conjugated characteristics may have abnormally large σ and η . Example of these outliers can be found in Figure 4.4. Therefore, these selective bonds, as stated in the list under category **D** above, are retained during the bond elimination step.

The remaining bonds are used to generate a list of dihedral angles, by selecting additional atoms, preferring heteroatoms to carbon and carbons to hydrogen.

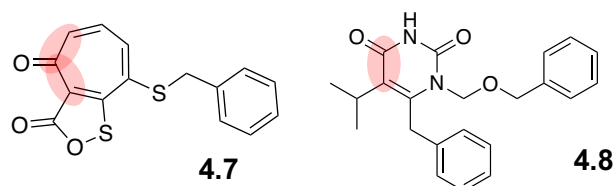
Odd cases type 1: conjugated bonds with heteroatom(s), e.g. C-N, C-S, N-S (typically, bond order = 1.5)

Examples:



Odd cases type 2: C-C(=O/N) bonds in conjugated systems (bond order = 1.5)

Examples:



Odd cases type 3: -C=N(non-ring atom) bonds (bond order = 2)

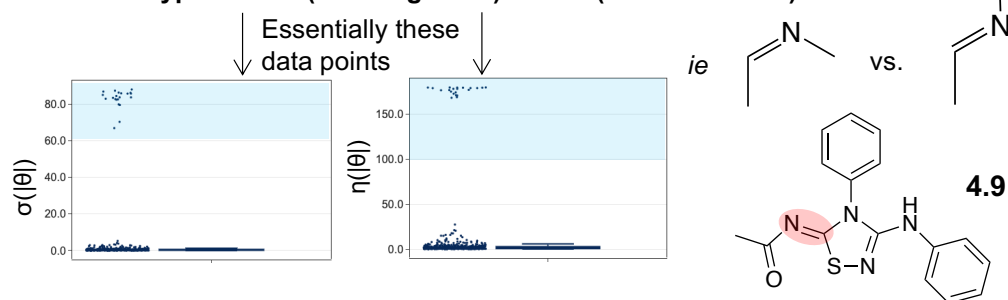


Figure 4.4. Inspecting the odd cases: examples and descriptions. $\eta(\theta)$ and $\sigma(\theta)$ values, or $\eta(\theta)$, $\sigma(|\theta|)$ and $\eta(|\theta|)$ for trans-oriented dihedrals, are abnormally large or small for the odd cases. These data points are identified as outliers in box plots. Examples of molecules that contain these bonds are given above.

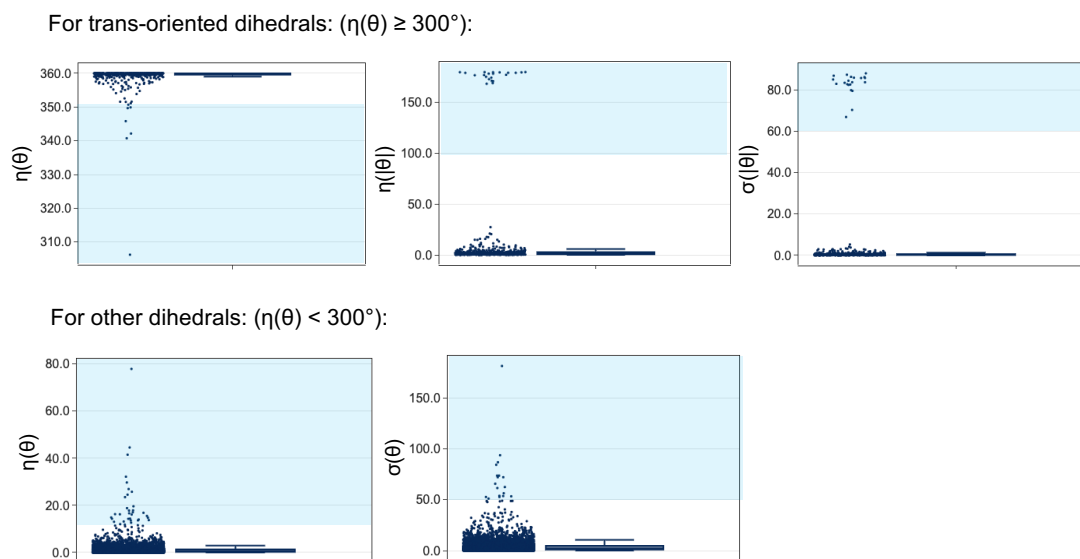


Figure 4.5. Identifying odd cases with box plots. Box plot analyses on standard deviation (σ) and range (η) parameters of dihedral angle values (θ) and absolute dihedral angle values ($|\theta|$). The box plots are based on 9017 sets of dihedral values. Dihedrals with a $\eta(\theta) \geq 300^\circ$ are identified as trans-oriented dihedrals. The outliers for inspections are in the blue region. Specifically, trans-oriented dihedrals with $\eta(\theta) < 350^\circ$, $\sigma(|\theta|) > 60^\circ$ or $\eta(|\theta|) > 100^\circ$ are outlier. For other dihedrals, the dihedrals are identified outlier if $\eta(\theta) > 15^\circ$ or $\sigma(\theta) > 50^\circ$.

The elimination criteria do not cover all possible fixed dihedrals. For example, dihedrals within the fused ring and most of dihedrals with conjugated C-X (X = heteroatoms) bonds are fixed. For each dihedral in the list of potentially rotatable dihedrals, dihedral angle values (θ) are calculated for every conformer. The standard deviation and range parameter of these dihedral angle values ($\sigma(\theta)$ and $\eta(\theta)$) are also computed. For dihedrals with values of θ close to $\pm 180^\circ$, the standard deviation and range are computed from absolute θ (i.e. $\sigma(|\theta|)$ and $\eta(|\theta|)$). If these parameters are within the limits given in Table 4.1, the dihedral is then labelled as fixed and the associated dihedral angle data is not used as a descriptor.

Table 4.1. Criteria for identifying fixed dihedrals based on standard deviation (σ) and range (η) parameters of dihedral angle values (θ) and absolute dihedral angle values ($|\theta|$). The parameters were derived from box plot analyses on 7262 sets of dihedrals, which satisfy the category **D**. Dihedrals with a $\eta(\theta) \geq 300^\circ$ are identified as trans-oriented dihedrals.

$\eta(\theta)$	$\eta(\theta)$	$\sigma(\theta)$	$\eta(\theta)$	$\sigma(\theta)$
$< 300^\circ$	$< 8.7^\circ$	$< 2.4^\circ$	-	-
$\geq 300^\circ$	$> 359.2^\circ$	-	$< 6.6^\circ$	$< 1.5^\circ$

The parameters in Table 4.1 were derived from box plot analyses on 7262 sets of dihedrals, which satisfy the category **D**. Box plot analyses were then performed on σ and r parameters of θ and $|\theta|$ values for these 7262 dihedrals (Figure 4.6). The upper fence or lower fence values from the box plots were taken as the parameter for filtering fixed bond in Table 4.1.

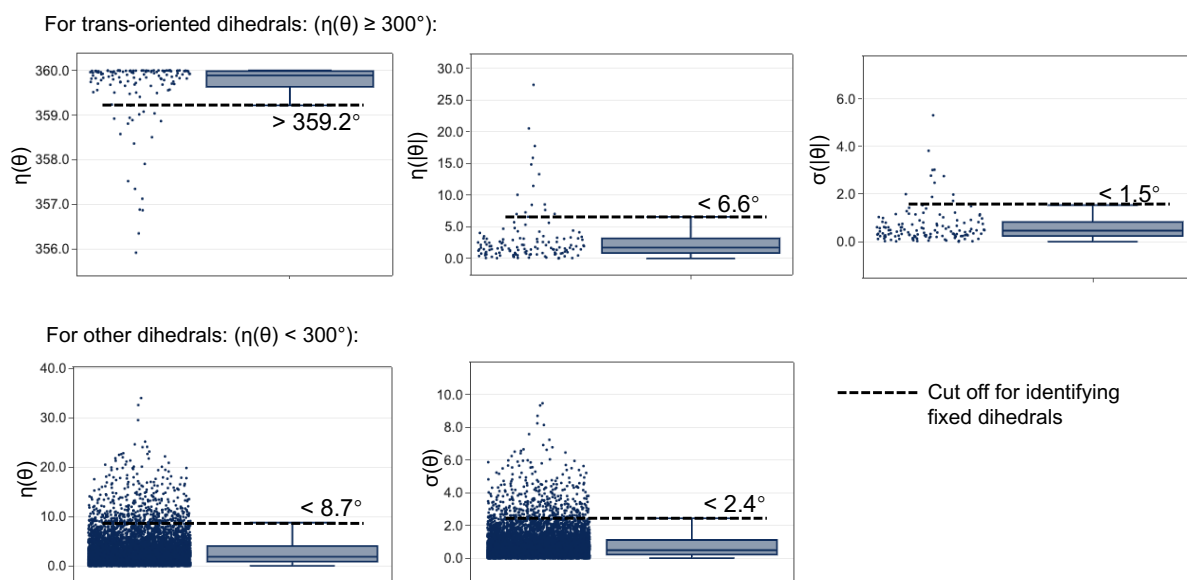


Figure 4.6. Deriving the criteria for identifying fixed bonds in Table 4.1: Box plot analyses on standard deviation (σ) and range (η) parameters of dihedral angle values (θ) and absolute dihedral angle values ($|\theta|$). The box plots are based on 7262 sets of dihedral values.

Figure 4.7 presents outcomes of the dihedral selection for an example molecule, **4.10**, following the above procedure.

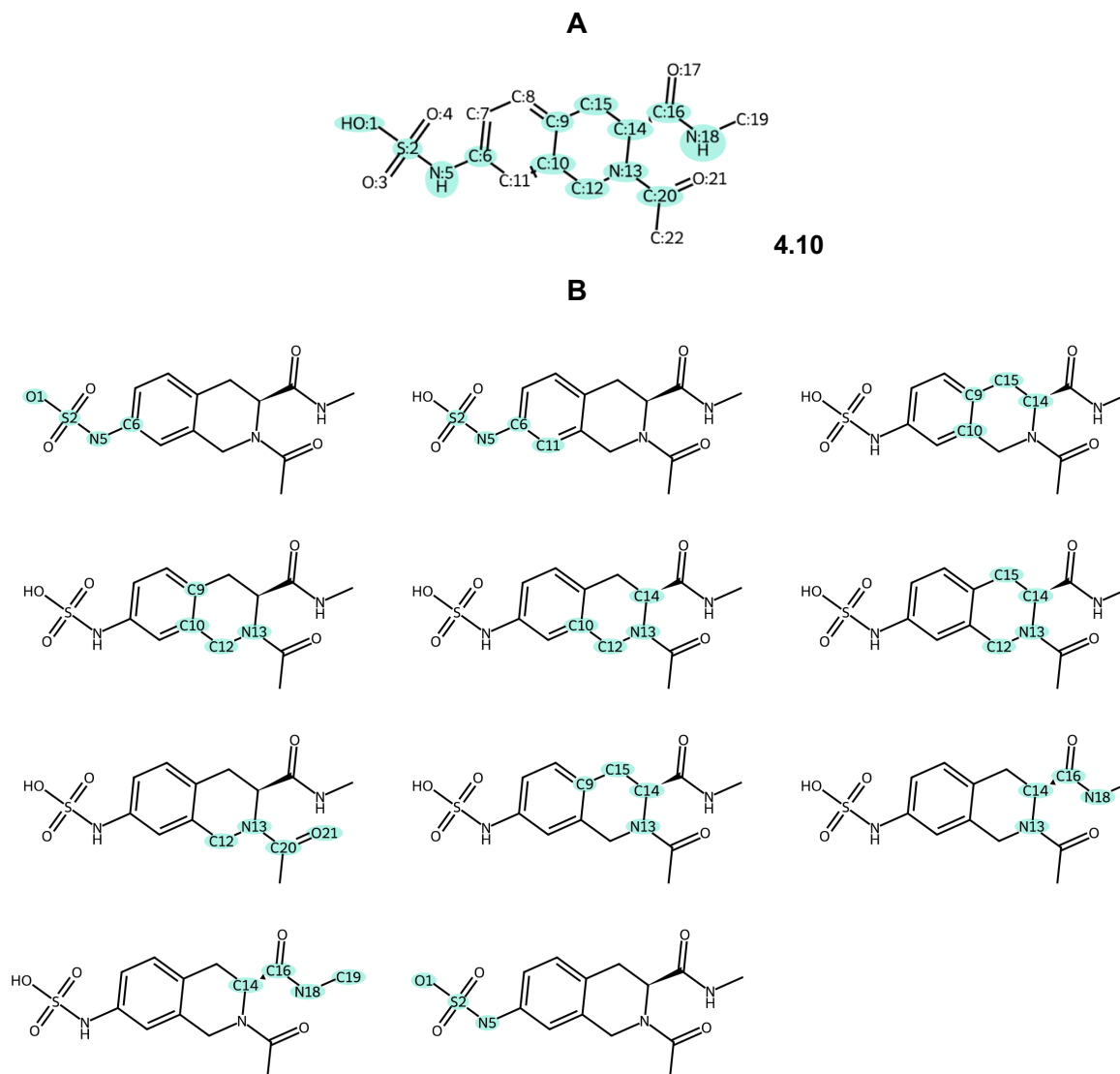


Figure 4.7. The dihedral selection for an example molecule, **10**. A) The example: the central atoms of the selected dihedrals are highlighted in blue. B) A breakdown of the selected dihedrals for the example molecule: all the non-H atoms in the selected dihedral are labelled and highlighted in blue.

Lastly, a post-processing procedure is introduced to remove the artefact of θ values as an 1D measurement. A 179° and a -179° dihedral may be numerically different but are structurally similar. In these cases, the θ values from the negative quadrant are shifted to join up with the data points in the positive quadrant (Figure 4.8). The process ensures the continuity of θ values within the same cluster.

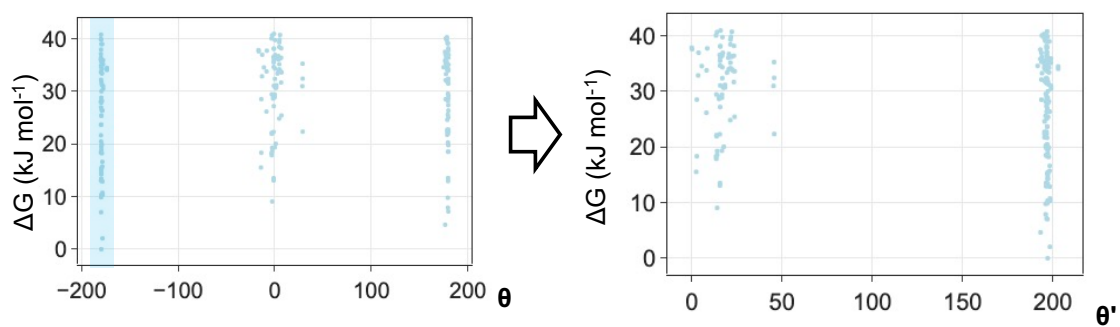


Figure 4.8. Post-processing on dihedral angle values (θ). The shifted data points are highlighted in blue.

The output of the above analyses and calculations is an array of dihedral angle data, ready for the clustering process.

4.3.2 Clustering algorithms

The clustering process is performed with the scikit-learn packages.¹³² The pipeline was initially developed using the K-means clustering algorithm. After finalising other parts of the pipeline, a benchmarking test was performed. Several clustering algorithms were tested using the default hyperparameters, including mini batch K-means clustering, agglomerative clustering, BIRCH and Gaussian mixture. Overall, changes in the clustering algorithms do not alter the performance significantly (Appendix Figure S4.2). However, the time required is noticeably different between clustering methods. The pipeline using the agglomerative clustering algorithm requires the least computational resource. The average time from extracting descriptors to the end of the clustering process is approximately a second (Figure 4.9). Hyperparameter tuning on the agglomerative clustering process was performed. Combinations of the linkage criterion and affinity metrics were tested in a grid searching style. Altering the hyperparameters does not change the result significantly (Appendix Figure S4.3). The agglomerative clustering method with the default hyperparameters (affinity = Euclidean and linkage = Ward) was chosen for the clustering approach pipeline. The inputs required are the list of dihedral angles, and *n_clusters*, the number of clusters that is required, which are discussed in the next section.

Method	Total calculation time (150 molecules)	Time per molecule
K-Means	01:59:11	00:00:48
BIRCH	00:07:55	00:00:03
Agglomerative Clustering	00:02:18	00:00:01
MiniBatchKMeans	00:52:51	00:00:21
GaussianMixture	00:42:12	00:00:17

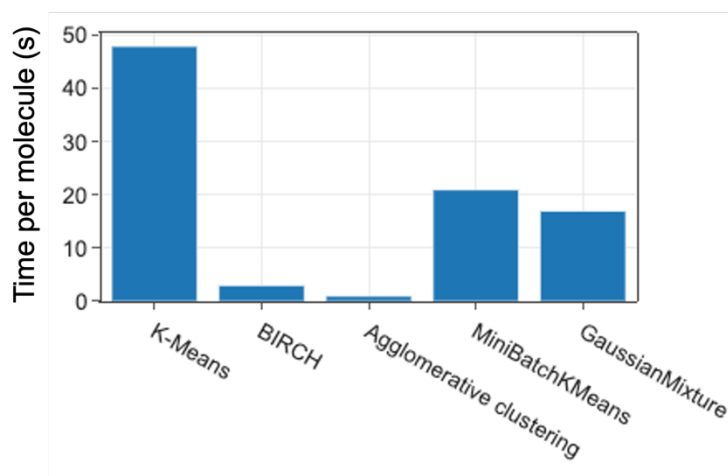


Figure 4.9. The benchmarking test on clustering algorithms: computational time. The total calculation time and time per molecule (or SDF) are reported above.

4.3.3 Priority list generation methods

This section describes the process of selecting conformations from the FF-re-optimised list for re-optimisation by DFT methods and ordering them into a priority list for re-optimisation. Three different approaches were used: *pipeline-x*; *pipeline-ascent* and *pipeline-mix* (Figure 4.10).

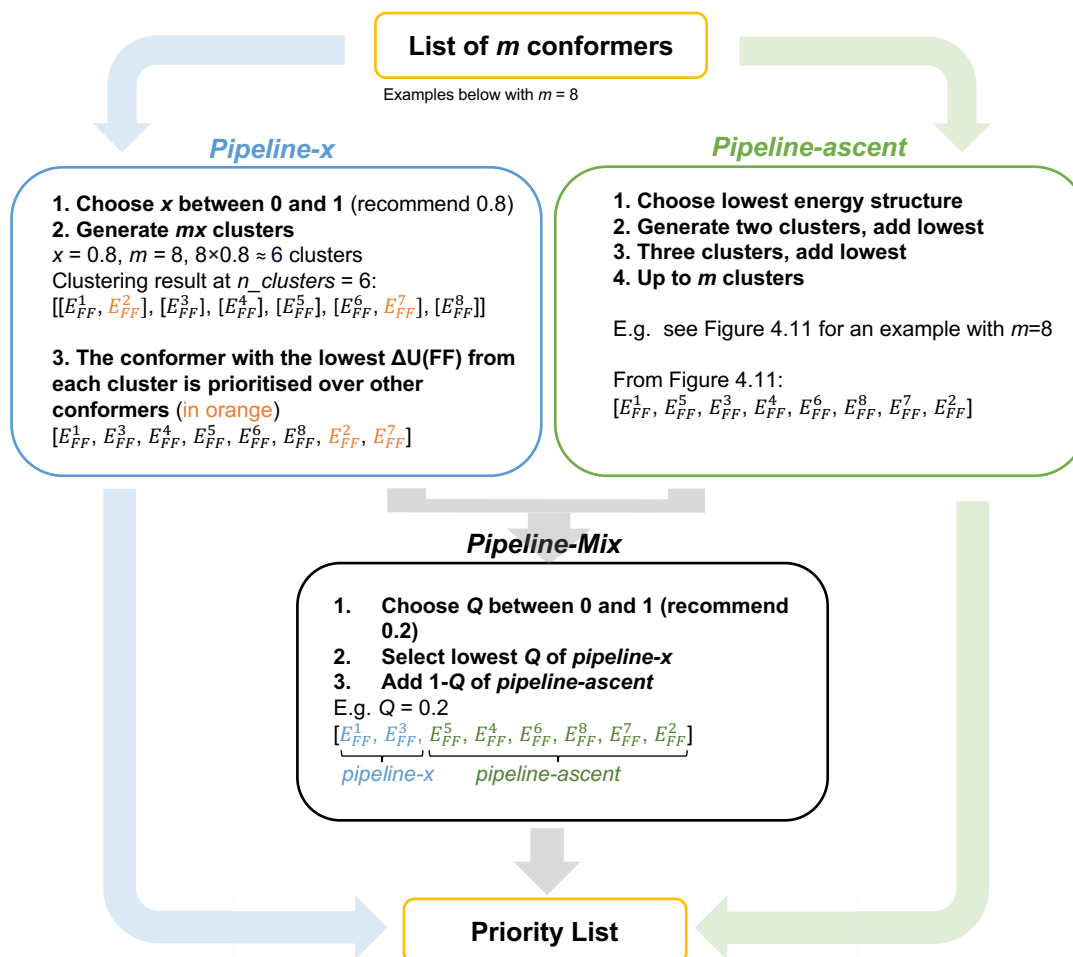


Figure 4.10. The different methods for generating priority lists from the clustering results. m is the total number of conformers in conformational searching output file.

The conformer with the lowest $\Delta U(FF)$ from each cluster is prioritised over other conformers. For the first of these, *pipeline-x*, the $n_clusters$ value is chosen as a fraction, x , of the number of conformers. Too small a value for x would group diverse structures together; too large a value would ignore similarities between structures. A good choice of x might lead to a distinct cluster for each unique DFT conformation. Since 85 % of the re-optimised structures in the dataset are unique, a slightly smaller value, 80 % of the total number of conformers, was selected as a good choice for $n_clusters$ ($x = 0.8$).

$n_clusters$	List of conformer clusters	Priority list
1	$[[E_{FF}^1, E_{FF}^2, E_{FF}^3, E_{FF}^4, E_{FF}^5, E_{FF}^6, E_{FF}^7, E_{FF}^8]]$	$[E_{FF}^1]$
2	$[[E_{FF}^1, E_{FF}^2, E_{FF}^3, E_{FF}^6, E_{FF}^7, E_{FF}^8], [E_{FF}^5, E_{FF}^4]]$	$[E_{FF}^1, E_{FF}^5]$
3	$[[E_{FF}^1, E_{FF}^2], [E_{FF}^3, E_{FF}^4], [E_{FF}^5, E_{FF}^6, E_{FF}^7, E_{FF}^8]]$	$[E_{FF}^1, E_{FF}^5, E_{FF}^3]$
4	$[[E_{FF}^1, E_{FF}^2], [E_{FF}^3], [E_{FF}^4], [E_{FF}^5, E_{FF}^6, E_{FF}^7, E_{FF}^8]]$	$[E_{FF}^1, E_{FF}^5, E_{FF}^3, E_{FF}^4]$
5	$[[E_{FF}^1, E_{FF}^2, E_{FF}^5], [E_{FF}^3], [E_{FF}^4], [E_{FF}^6, E_{FF}^7], [E_{FF}^8]]$	$[E_{FF}^1, E_{FF}^5, E_{FF}^3, E_{FF}^4, E_{FF}^6, E_{FF}^8]$
6	$[[E_{FF}^1, E_{FF}^2], [E_{FF}^3], [E_{FF}^4], [E_{FF}^5], [E_{FF}^6, E_{FF}^7], [E_{FF}^8]]$	$[E_{FF}^1, E_{FF}^5, E_{FF}^3, E_{FF}^4, E_{FF}^6, E_{FF}^8]$
7	$[[E_{FF}^1, E_{FF}^2], [E_{FF}^3], [E_{FF}^4], [E_{FF}^5], [E_{FF}^6], [E_{FF}^7], [E_{FF}^8]]$	$[E_{FF}^1, E_{FF}^5, E_{FF}^3, E_{FF}^4, E_{FF}^6, E_{FF}^8, E_{FF}^7]$
8	$[[E_{FF}^1], [E_{FF}^2], [E_{FF}^3], [E_{FF}^4], [E_{FF}^5], [E_{FF}^6], [E_{FF}^7], [E_{FF}^8]]$	$[E_{FF}^1, E_{FF}^5, E_{FF}^3, E_{FF}^4, E_{FF}^6, E_{FF}^8, E_{FF}^7, E_{FF}^2]$

Figure 4.11. Illustrating the *pipeline-ascent* priority list generating method for a molecule with eight conformers: from the clustering result (ie lists of conformer clusters) to the final priority list. At each new value of $n_clusters$, the lowest energy structures are added from every cluster without a member on the priority list. The blue colouring indicates the lowest energy conformer from each cluster by FF energy in the list of conformer clusters. The green colouring indicates a conformer that has not appeared before and is added to the priority list at the corresponding $n_clusters$ value.

For the *pipeline-ascent* method (Figure 4.11), the clustering process is repeated with all possible values of $n_clusters$ from one to the total number of conformations (m), which gives m list of conformer clusters. From clustering results with $n_clusters = 1$ to $n_clusters = m$, the lowest energy conformers from every cluster are introduced into the priority list while omitting those that are already in the priority list.

The *pipeline-mix* method involves truncating the *pipeline-x* and *pipeline-ascent* priority list. The Q parameter is the ratio of the *pipeline-x* content in the new *pipeline-mix* priority list. A new priority is assembled by picking the lowest Q of the *pipeline-x* priority list. Then the list is completed in order of the *pipeline-ascent* priority list while omitting conformers that have already been chosen.

In this thesis, the clustering approach with the *pipeline-x* method is referred as the *pipeline-x* approach. A similar logic applies to the *pipeline-ascent* and *pipeline-mix* approach.

The above sections describe the procedure of CONFPASS in generating the priority list for conformer re-optimisations at a higher level of theory.

4.4 Evaluating CONFPASS priority lists

4.4.1 Evaluation methods

4.4.1.1. The DFT list

The DFT dataset, where all FF structures were re-optimised at the DFT level, was utilised to test and investigate how close CONFPASS gets to the complete analysis after re-optimising only a fraction of the FF structures. FF structures that lead to duplicate conformers at the DFT level should be least prioritised.

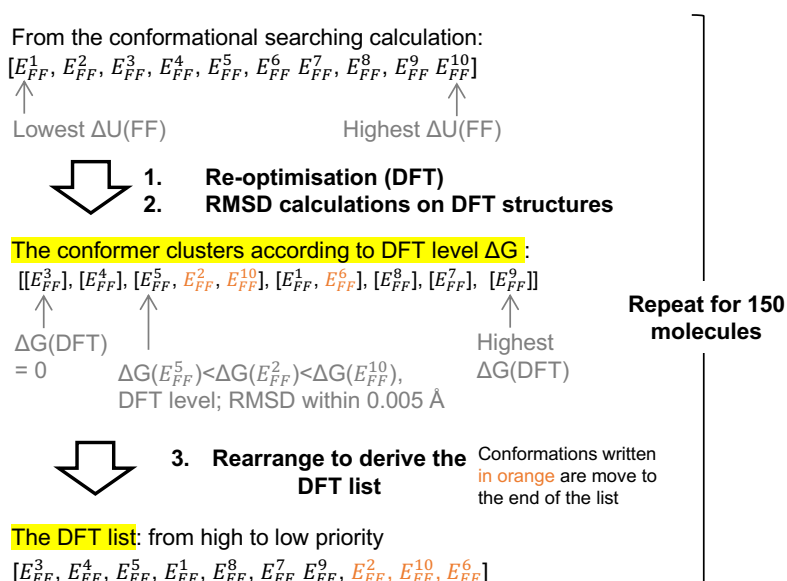


Figure 4.12. The process flowchart for obtaining the DFT list for evaluation. In the above example, conformer E_{FF}^5 , E_{FF}^2 , and E_{FF}^{10} have the same structure after re-optimisation at DFT level with root-mean-square deviation (RMSD) values less than 0.005 Å. Conformer E_{FF}^5 has the lowest $\Delta G(DFT)$ among the three and thus is prioritised over conformer E_{FF}^2 and E_{FF}^{10} . The same logic applies to the $[E_{FF}^1, E_{FF}^6]$ cluster.

A workflow was built and used to derive the DFT list from the conformational searching output of a molecule (Figure 4.12). The conformers are given an index according to the internal energy at the FF level ($\Delta U(FF)$). Firstly, DFT optimisations were conducted on all conformers in accordance with the DFT dataset described in Chapter 4.2 Data Gathering. Root-mean-square deviation (RMSD) calculations were performed with the GetBestRMS function under the rdkit.Chem.rdMolAlign module to compare the DFT structure on all non-H atoms of each pair of conformers.⁴² If the RMSD value is less than 0.005 Å, the pair of conformers would be considered as having the same structure. DFT conformers with the same structure were grouped into the same clusters. The clusters and the conformers within the clusters were arranged based on their ΔG values at the DFT level ($\Delta G(DFT)$), which led to conformer clusters at the DFT level. The range of $\Delta G(DFT)$ within a cluster is usually less than 0.1 kcal mol⁻¹. The DFT lists were constructed from conformer clusters at the DFT level. In the DFT lists, conformers with the lowest

$\Delta G(\text{DFT})$ within each cluster were prioritised over others. The above process was repeated for 150 molecules in the DFT dataset.

Proposed priority lists from FF structures were compared to the DFT list and conformer clusters at the DFT level, for evaluation. The global minimum test and bins test were established and conducted to measure how well priority generation approaches prioritise the lowest energy conformers and unique conformer structures, respectively.

4.4.1.2. The global minimum measure

The ratio of the number of DFT re-optimised conformers to the total number of conformers at FF level is called r_{opt} . CONFPASS tries to minimise the value of r_{opt} which corresponds to finding the global minimum. This r_{opt} value is P_{GMT} (Figure 4.13).

A proposed priority list:

$$\underbrace{[E_{\text{FF}}^1, E_{\text{FF}}^2, E_{\text{FF}}^3, E_{\text{FF}}^4, E_{\text{FF}}^5, E_{\text{FF}}^6, E_{\text{FF}}^7]}_{\substack{\text{Not the lowest energy conformer but} \\ \text{have higher priority in the priority list}}} \quad P_{\text{GMT}} = \frac{3}{7} = 0.43$$

The lowest energy conformer after DFT reoptimisations: E_{FF}^3

Figure 4.13. Tests for evaluating the performance of priority list generation approaches in prioritising the lowest energy conformer, the global minimum test.

4.4.1.3. The bins test

Each distinct structure at the DFT level is put in a separate ‘bin’. The bins test, Δa_{bins} , is a measure of how well the priority list avoids prioritising duplicate conformers. The 150 molecules in the DFT dataset have 15 342 conformers at FF level. In the 15 342 re-optimisation calculations at the DFT level, 2280 (or 14.9%) did not lead to a unique structure. There are 13 062 unique structures, thus 13 062 bins, in the complete dataset.

The number of populated bins for a particular selection of conformers from partial optimisations divided by the total number of distinct structures at the DFT level is labelled r_{bins} (Figure 4.14.A). A plot of r_{bins} vs r_{opt} plot (Figure 4.14.B) shows how r_{bins} increases with r_{opt} , the fraction of the total number of structures that has been re-optimised. For the optimal order, r_{bins} should increase linearly with r_{opt} and plateau at $r_{\text{bins}} = 1$. The plot of r_{bins} vs r_{opt} for a proposed priority list is always under the line of the DFT list. The bins test, Δa_{bins} , is the area between the optimal order and the actual result in a r_{bins} vs r_{opt} plot. A proposed priority list with a small Δa_{bins} value is desirable as conformers with unique structures are prioritised and duplicate conformers are avoided like in the DFT list. CONFPASS tries to minimise Δa_{bins} .

A

Conformer clusters at the DFT level:

$[[E_{FF}^3], [E_{FF}^4], [E_{FF}^5, E_{FF}^2, E_{FF}^{10}], [E_{FF}^1, E_{FF}^6], [E_{FF}^8], [E_{FF}^7], [E_{FF}^9]] \rightarrow 7 \text{ bins}$

The DFT priority list:

$[E_{FF}^3, E_{FF}^4, E_{FF}^5, E_{FF}^1, E_{FF}^8, E_{FF}^7, E_{FF}^9, E_{FF}^2, E_{FF}^{10}, E_{FF}^6]$

r_{opt}	Optimised conformers	Conformer clusters at the DFT level	n_{bins}	r_{bins}
1.0	$[E_{FF}^3, E_{FF}^4, E_{FF}^5, E_{FF}^1, E_{FF}^8, E_{FF}^7, E_{FF}^9, E_{FF}^2, E_{FF}^{10}, E_{FF}^6]$	$[[E_{FF}^3], [E_{FF}^4], [E_{FF}^5, E_{FF}^2, E_{FF}^{10}], [E_{FF}^1, E_{FF}^6], [E_{FF}^8], [E_{FF}^7], [E_{FF}^9]]$	7	7/7
0.8	$[E_{FF}^3, E_{FF}^4, E_{FF}^5, E_{FF}^1, E_{FF}^8, E_{FF}^7, E_{FF}^9, E_{FF}^2, E_{FF}^{10}]$	$[[E_{FF}^3], [E_{FF}^4], [E_{FF}^5, E_{FF}^2], [E_{FF}^1], [E_{FF}^8], [E_{FF}^7], [E_{FF}^9]]$	7	7/7
0.6	$[E_{FF}^3, E_{FF}^4, E_{FF}^5, E_{FF}^1, E_{FF}^8, E_{FF}^7]$	$[[E_{FF}^3], [E_{FF}^4], [E_{FF}^5], [E_{FF}^1], [E_{FF}^8], [E_{FF}^7]]$	6	6/7
0.4	$[E_{FF}^3, E_{FF}^4, E_{FF}^5, E_{FF}^1]$	$[[E_{FF}^3], [E_{FF}^4], [E_{FF}^5], [E_{FF}^1]]$	4	4/7
0.2	$[E_{FF}^3, E_{FF}^4]$	$[[E_{FF}^3], [E_{FF}^4]]$	2	2/7

n_{bins} = the number of populated bins for a particular selection of conformers from partial optimisations
 $r_{\text{bins}} = n_{\text{bins}} / \text{total number of bins in the conformer clusters at DFT level}$

B

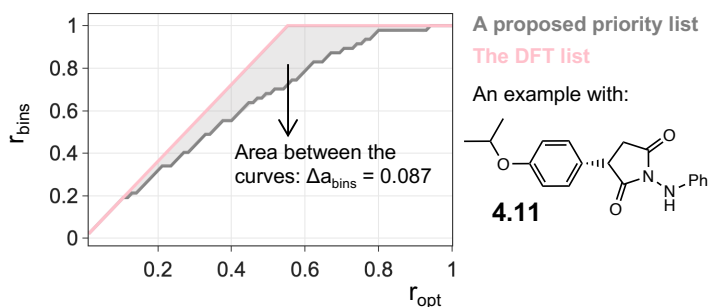


Figure 4.14. Tests for evaluating the performance of priority list generation approaches in prioritising unique conformer structures, the bins test. A) The derivation of r_{bins} values. r_{bins} is the ratio of populated bins for a particular selection of conformers from partial optimisations. n_{bins} is the number of populated bins in the conformer cluster at the DFT level. The orange colouring indicates duplicate conformers at DFT level. B) An example r_{bins} vs r_{opt} plot.

4.4.1.4. Overall test

The overall parameter (P_{overall} , eq. 2) accounts for both the global minimum test (P_{GMT}) and bins test (Δa_{bins}) parameters.

$$P_{\text{overall}} = P_{\text{GMT}} + 5 \Delta a_{\text{bins}} \quad (2)$$

Theoretically, Δa_{bins} can take up a value from 0 to 1. Preliminary studies show that 98.6% of the Δa_{bins} values sampled are less than 0.2 (Figure 4.15), which is also in line with percentage of duplicate structures in the DFT dataset, 14.9%. Hence, Δa_{bins} is scaled by a factor of five. A low overall parameter value is desirable as this implies a good performance of the proposed priority list in prioritising the global minimum and unique conformations.

Hyperparameter tuning tests were conducted to validate the choice of default x and Q value with P_{overall} . Analyses (See Figure 4.16) showed the best performance comes approximately when $x = 0.8$ and $Q = 0.2$ for the *pipeline-x* and *pipeline-mix* approach.

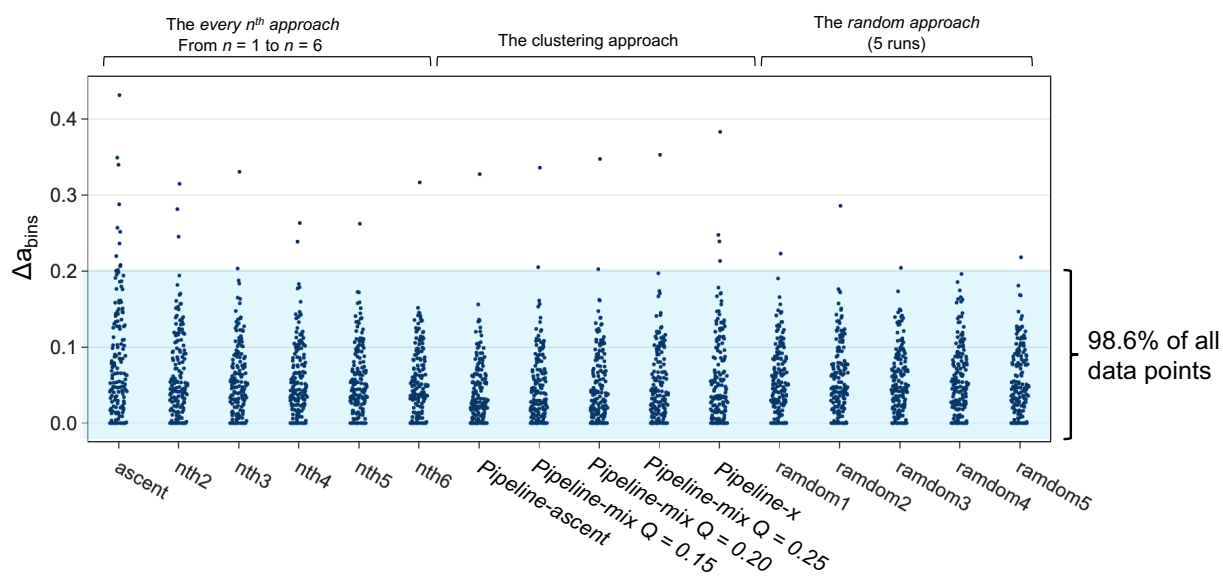


Figure 4.15. Distributions of Δa_{bins} values for the 150 molecules with different priority list generation approaches. 98.6% of data points have a value less than 0.2 (ie the light blue region).

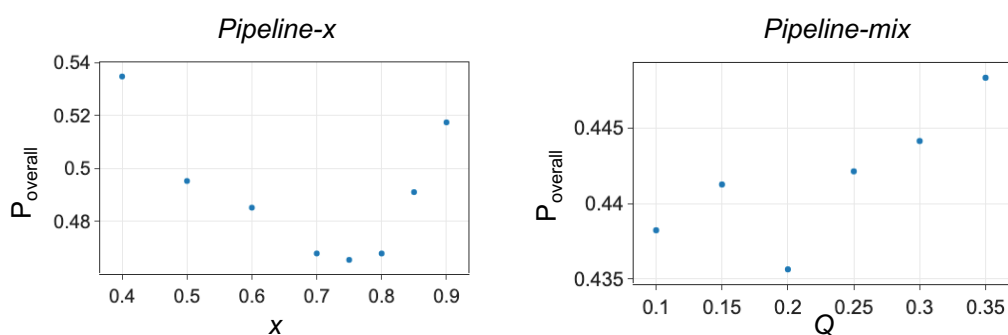


Figure 4.16. Parameter tuning for the *pipeline-x* and *pipeline-mix* approach: the overall parameter (P_{overall}) vs x and overall parameter vs Q plot.

4.4.2 Results

A standard procedure has been followed to compare the effectiveness of different priority list generation approaches (Figure 4.17). Firstly, a priority list was generated from analyses of the conformational searching output file with non-pipeline or pipeline approaches. The proposed priority list was compared with the DFT list via the global minimum test and bins test, where P_{GMT} and Δa_{bins} were calculated. The above process was repeated for the 150 molecules in the DFT dataset. A pair of (mean Δa_{bins} , mean P_{GMT}) can be obtained for every priority list generation approach.

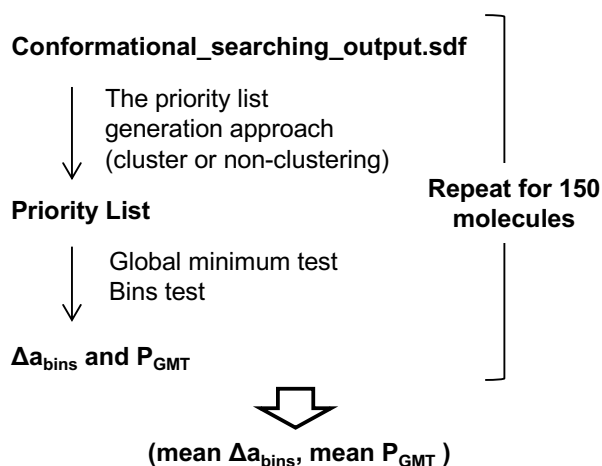


Figure 4.17. The workflow for evaluating pipeline and non-pipeline priority list generation approaches using the DFT dataset.

The mean P_{GMT} vs mean Δa_{bins} plot illustrates the difference in performance between pipeline and non-pipeline approaches (Figure 4.18).

In the non-pipeline approaches, the *random approach* gives a lower mean Δa_{bins} and performs better in avoiding prioritising duplicate conformers compared to the *every n^{th} approach*. However, the *every n^{th} approach*, particularly the *ascent* approach ($n = 1$), gives a better chance of obtaining the global minimum with fewer re-optimisations. Increasing the n value reduces the global minimum test performance but improves the bins test performance.

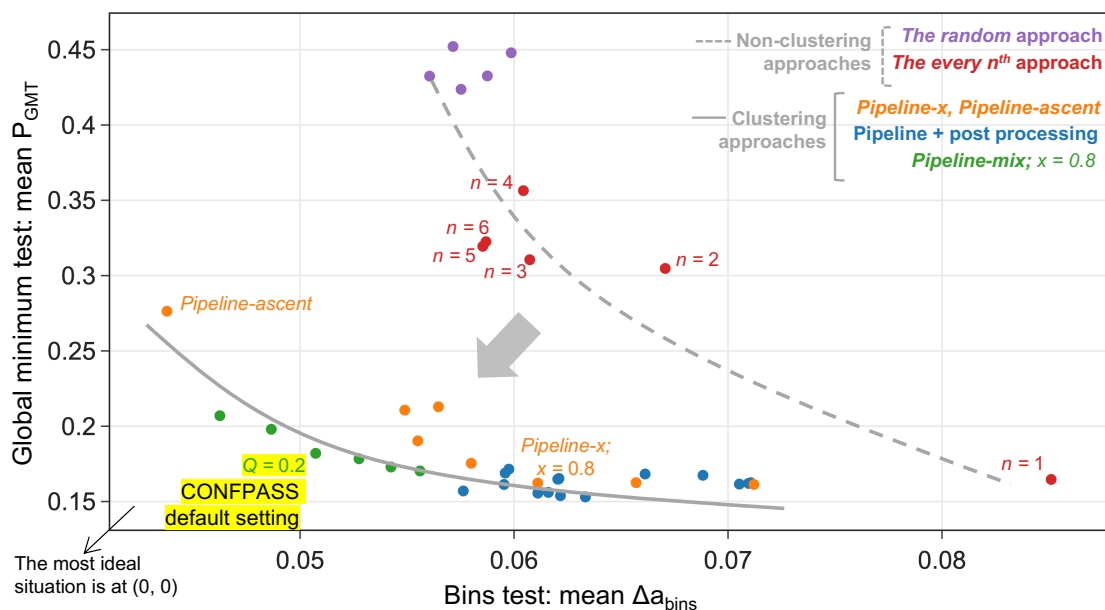


Figure 4.18. The global minimum test (P_{GMT}) and bins test (Δa_{bins}) results are presented together in a scatter plot. The tests for the *random approach* have been repeated for 5 times, which contributes to the 5 data points under this category. The corresponding hyperparameters (*ie* n , x or Q) are labelled for selective approaches.

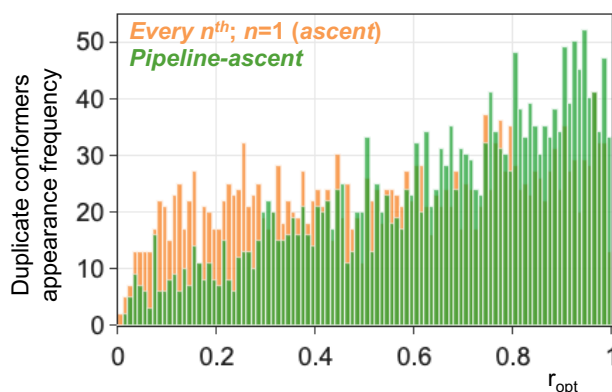


Figure 4.19. The histogram of duplicate conformer appearance frequency over r_{opt} . The *pipeline-ascent* approach arranges the duplicate conformers toward the back of priority list.

In the pipeline approach, the *pipeline-ascent* approach has the best bins test performance compared to all other approaches. $\Delta a_{\text{bins}}(\text{pipeline-ascent})$ is almost halved compared to $\Delta a_{\text{bins}}(\text{ascent})$. Additional histogram analyses of duplicate conformer appearance frequency over the re-optimisation process in terms of r_{opt} have been conducted (Figure 4.19). The chance of getting duplicate conformers with the *pipeline-ascent* priority lists is approximately halved during the first 30% of the re-optimisations compared to the *ascent* approach. The mean P_{GMT} of the *pipeline-x* approach is similar to the *ascent* approach. Mean $P_{\text{GMT}} = 0.14$ implies that approximately 14% of the conformers in the priority need to be re-optimised before obtaining the global minimum structure. However, the mean Δa_{bins} value of the

pipeline-x approach is noticeably lower than the *pipeline-ascent* approach. The *pipeline-mix* approach, with the parameter Q set to 0.2, improves the bins test parameter without significantly reducing the global minimum test performance. With this setting, the global minimum structure can be obtained in the first 30% of the re-optimisations about 80% of the time (Figure 4.20).

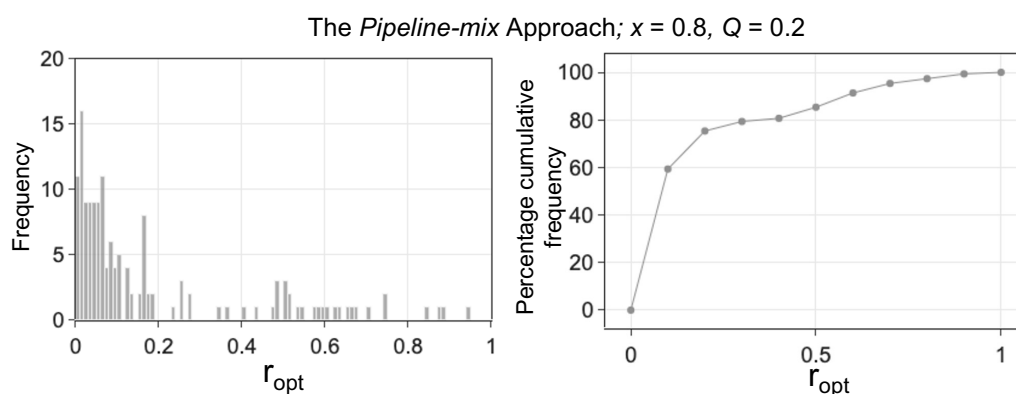


Figure 4.20. Histogram analyses on the distribution of global minimum test measurement, P_{GMT} , distribution with the *pipeline-mix* approach.

Further investigations on whether further steps improve the performance were performed (details in Appendix to Chapter 4, 4.C). The aim is to extract and use the energy and structural data from the re-optimised conformers to improve the proposed priority list. The ideas that have been attempted include:

- 1) Sequential clustering process with certain FF structures replaced by the re-optimised DFT structures in the input SDF;
- 2) Varying the x parameter according to percentage of structural duplicates in the group of re-optimised conformers when using the *pipeline-x* method;
- 3) Prioritising conformers from the same clusters as the lowest energy DFT re-optimised structure;
- 4) Using clusters-distance array from the K-Means clustering process to make the prioritisation decisions.

Post-processing procedures are not applicable for rigid molecules with less than 3 conformers. The steps involving sequential clustering processes require further developments for radical species if their pseudo structure and actual structure have different atom numbers. For the post-processing procedures, only 147 molecules from the DFT dataset were considered during evaluation. Overall, only marginal improvements in the global minimum test were observed (Figure 4.18, marked as post-processing). This demonstrates the effectiveness of the standard pipeline approaches.

Implementing the pipeline approach in generating the priority list led to a shift in data points toward the bottom left corner in the mean P_{GMT} vs mean Δa_{bins} plot. The clustering process with dihedral angles as descriptors offered by the pipeline effectively groups conformationally similar conformers, which led to the noticeable improvement in the bins test parameter by the pipeline approaches. Differences in performance are also reflected in the mean overall parameter (Figure 4.21), calculated from the mean

Δa_{bins} and mean P_{GMT} . The overall parameters for all other pipeline approaches are noticeably lower compared to the non-clustering approaches. Overall, the clustering approaches outperform the non-pipeline approach in prioritising the global minimum structure and without duplications from different force field (FF) structures re-optimising to same structure.

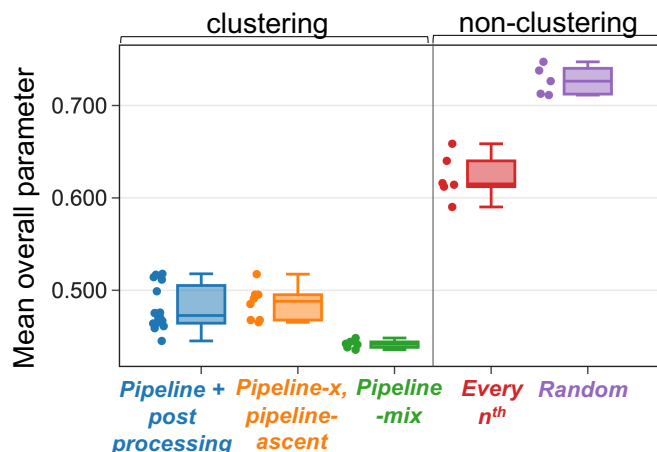


Figure 4.21. The overall parameter results for the pipeline and non-pipeline approaches are compared with boxplot. The tests for the *random* approach have been repeated for 5 times, which contributes to the 5 data points under this category. For the clustering approach, various different hyperparameter settings have been considered.

If the aim of the study is to find the conformer with the global minimum energy at DFT level, *pipeline-x* is the best approach. On the other hand, the *pipeline-ascent* approach should be used if only structural diversity is needed. This approach receives the lowest Δa_{bins} in the bins test. Finally, the *pipeline-mix* approach should be used if the user wishes to avoid re-optimising duplicate structures and want to obtain global minimum structure with less computational resources. The *pipeline-mix* approach with $x = 0.8$ and $Q = 0.2$ outperforms all other approaches in terms of the overall parameter and is chosen as the default setting for the clustering approach pipeline in CONFPASS.

The pipeline approach development, test and evaluations are based on the MMFF to ω B97X-D/6-311G(d,p)//B3LYP-D3/6-31G(d) structural and energy data for 150 molecules. DFT datasets at other levels of theory are generated to test the generality of the clustering approach pipeline. Grayson *et al.* have conducted DFT re-optimisations with M06-2X functionals with implicit solvent model (IEFPCM/benzene) on all the conformers on 20 organic molecules from conformational searching calculations with various force fields (incl. MMFF, MM2 and OPLS3e).⁵² The data from the work of Grayson *et al.* was processed and formatted for the tests. The conformers of the 20 molecules were re-optimised at the expensive ω b97xd/6-311++g(d,p)// ω b97xd/6-31g(d) level of theory from the MMFF structures. In terms of the overall parameter, similar results were obtained when comparing the 4 subsets of 20 molecules and the extensive DFT dataset of 150 molecules (Appendix Figure S4.7). Additionally, the tests were repeated with $\Delta H(\text{DFT})$, $\Delta U(\text{DFT})$ and $\Delta U(\text{B3LYP-D3/6-31G(d)}/\text{B3LYP-}$

D3/6-31G(d)) values instead of $\Delta G(\text{DFT})$. Using alternative sets of energy did not change result patterns and the same conclusions can be drawn (Appendix Figure S4.10).

The recommended setting, also the default for CONFPASS, is the *pipeline-mix approach* with $x = 0.8$ and $Q = 0.2$. This usually gives a distribution of low energy conformations and also finds the global minimum conformation quickly.

4.5 Predicting the completion of the re-optimisation process

Every DFT re-optimisation gives us information about the conformation space of a molecule. Do we have enough information to decide we have gone far enough along the priority list and it is time to stop re-optimisations? To investigate this, the dataset of exhaustive DFT re-optimisations on FF conformation searches of 150 molecules (a total 15 342 FF conformers) was used to build a machine learning model for answering the question.

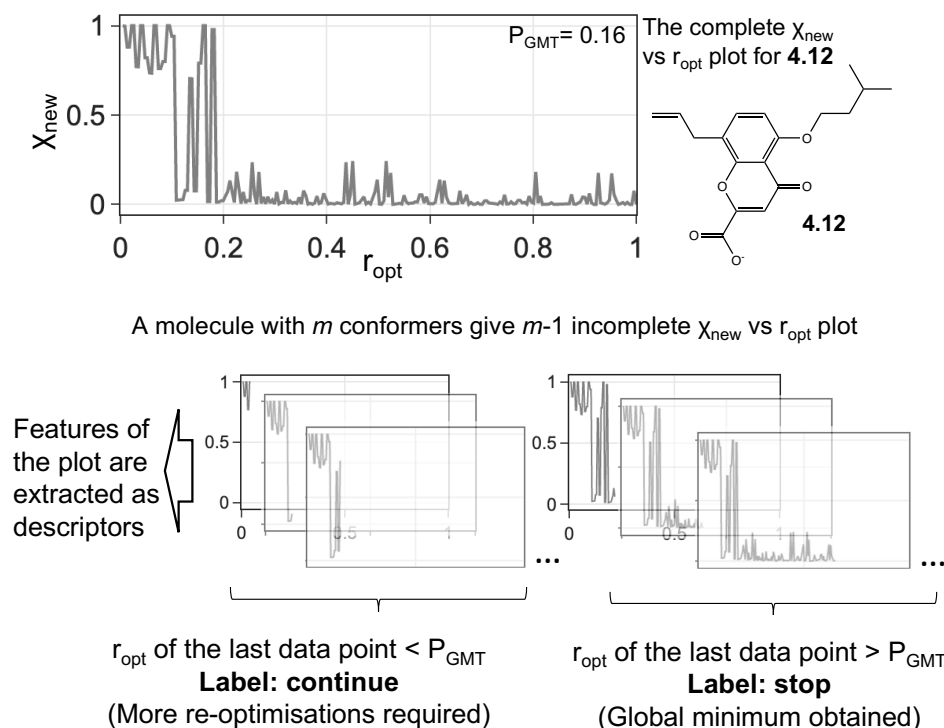


Figure 4.22. The mole amount of the last optimised conformer relative to the current global minimum (χ_{new}) vs ratio of optimised conformer (r_{opt}) plot. Each complete χ_{new} vs r_{opt} plot gives m sets of descriptor array and label, where m is the total number of conformers from conformational searches. P_{GMT} is the global minimum parameter.

As FF structures of a molecule are re-optimised following a priority list, the mole amount relative to the current global minimum of the new conformer, χ_{new} , can be calculated for each new re-optimised structure based on all conformers re-optimised so far (eq. 3; ΔG_{new} is the ΔG of the new conformer compared to the global minimum in the re-optimised structure so far).

$$\chi_{\text{new}} = \exp(-\Delta G_{\text{new}}/RT) \quad (3)$$

If the new structure is a duplicate of an earlier one, χ_{new} is set to zero. The proportion of the FF conformers that have been re-optimised, *ie* r_{opt} , is also known. After exhaustive DFT re-optimisations of every conformer, the χ_{new} is plotted against r_{opt} for the molecule (an example in Figure 4.22). The global minimum structure corresponds to the point with the highest r_{opt} value with $\chi_{\text{new}} = 1$ on the complete χ_{new} vs r_{opt} plot.

How do we decide when to stop? With an exhaustive χ_{new} vs r_{opt} plot, P_{GMT} , the value of r_{opt} at which the global minimum is found, can be determined. Usually, however, it is rare to re-optimize every conformer from the conformational searching output and so P_{GMT} is not known. Partial re-optimisations of the FF structures can provide an incomplete χ_{new} vs r_{opt} plot. Employing a machine learning approach can help to investigate if the data from an incomplete χ_{new} vs r_{opt} plot is sufficient to decide if the re-optimisation process structure can be terminated because the global minimum has probably been found.

The DFT dataset can provide the data for the model training. In addition to the full χ_{new} vs r_{opt} plot, each molecule should give $m-1$ incomplete χ_{new} vs r_{opt} plot, where m is the total number of conformers. The label (*ie* continue or stop) is assigned to each χ_{new} vs r_{opt} plot by comparing r_{opt} value of the last data point to the corresponding P_{GMT} . Features of a χ_{new} vs r_{opt} plot were extracted as descriptors:

1. Percentage of $\chi_{\text{new}} = 0$, percentage of $\chi_{\text{new}} = 1$, percentage of $\chi_{\text{new}} \leq 0.1$, percentage of $\chi_{\text{new}} \leq 0.2$ and percentage of $\chi_{\text{new}} \leq 0.5$ of the entire and the last 40% of the given χ_{new} vs r_{opt} plot
2. The highest r_{opt} value at which $\chi_{\text{new}} = 1$
3. r_{opt} value of the last data point in the plot
4. Total number of data points in the plot

The χ_{new} vs r_{opt} plot pattern depends on the order of the conformers in the priority list. Using priority lists generated under a specific setting, 150 molecules within the DFT dataset are expected to yield 15 342 descriptor arrays. Nine different priority list generation approaches have been considered to generate as many χ_{new} vs r_{opt} plots as possible for building up the training data: *pipeline-x* ($x = 0.8$), *pipeline-ascent*, three sets of *pipeline-mix* ($x = 0.8$; $Q = 0.15, 0.20, 0.25$) and four sets of *every n^{th}* ($n = 1-4$) priority list. This leads to $15\,342 \times 9 = 138\,078$ sets of a descriptor array and label. Given a list of χ_{new} and r_{opt} values for a partial re-optimisation, should the optimisation continue? The above gives the answer for the 138 078 examples from exhaustive re-optimisation. Can this training set be used to generate a reliable guide to how many structures in a new conformation search need to be re-optimised?

Various classification methods have been evaluated from scikit-learn¹³²: random forest (RF), K-nearest neighbour classification (KNN), logistic regression (LR) and Gaussian Naive Bayes (GaussianNB). To assess the performance of the models in predicting the completion of the re-optimisation process, a rigorous cross-validation methodology was followed. For each molecule, relevant data points associated with the molecule, which corresponds to less than 1% of the entire dataset, were taken out from the dataset. Models were trained on the filtered dataset. Input descriptor arrays derived from the default CONFPASS priority list generation approach (*pipeline-mix* with $x = 0.8$ and $Q = 0.2$) of the molecule were used for testing. Each molecule contributes m sets of prediction results from the designated model, where m is the total number of conformers. The above processes gave 15 342 prediction results for each classification method. The default hyperparameter settings were used. The LR model gives the highest accuracy, 0.88. The GaussianNB model has an accuracy of 0.86. KNN and RF model show similar performance with an accuracy of around 0.81. A local model from χ_{new} vs r_{opt} plots based on the *pipeline-mix* priority list ($x = 0.8$ and $Q = 0.2$) by itself was also built. The mean accuracy of the local model was very similar to the global model (Appendix Table S4.2).

The CONFPASS package is based on the global LR model. The LR model generates a probability value, p^{LR} , giving its confidence in each classification. Figure 4.23 presents a plot of p^{LR} vs r_{opt} from the cross-validation test of the global LR model. More than 50% of the false predictions have a r_{opt} value of less than 0.2. The outcome is within the expectation because the predictions rely on a descriptor array derived from a relatively limited amount of data.

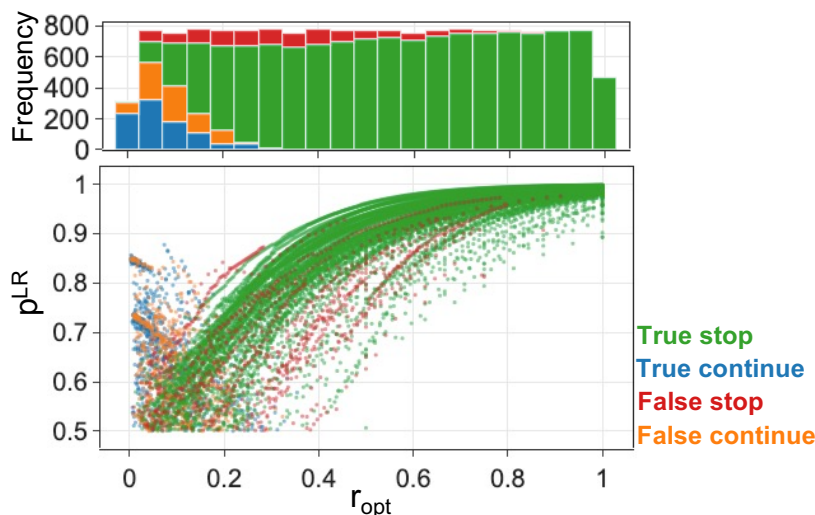


Figure 4.23. p^{LR} vs r_{opt} plot from the cross-validation test of the global LR model using input descriptor arrays derived from the default CONFPASS priority list generation approach. p^{LR} is the probability value from the LR model. A histogram is given for the true and false predictions by their r_{opt} values.

A cross-validation test was performed using the entire dataset (138 078 data points) with the default LR model and recorded down the probability value (p^{LR}) for every prediction. Histograms of p^{LR} for true and false predictions are given in Figure 4.24.A. Most true predictions have a p^{LR} value close to 1. The percentage of true prediction (%True) values within different p^{LR} ranges were calculated. The mean %True values were computed and plotted against the midpoint of the corresponding p^{LR} range.

percentage of true predictions can be used as a measurement for confidence, *ie* an equivalent to the percentage confidence (%Conf; mean %True = %Conf; %False = 1 – %True). Using the curve_fit function from scipy.optimize, a sigmoid function was fitted on the %Conf vs p^{LR} plot (eq. 4, Figure 4.24.B). The sigmoid function can be used to predict the %Conf value given the p^{LR} value of a prediction. For this model, $A = 6.29$. The mean absolute error of the fit was 2.6%.

$$\%Conf = \frac{1}{1 + e^{-A(p^{LR} - 0.5)}} \quad (4)$$

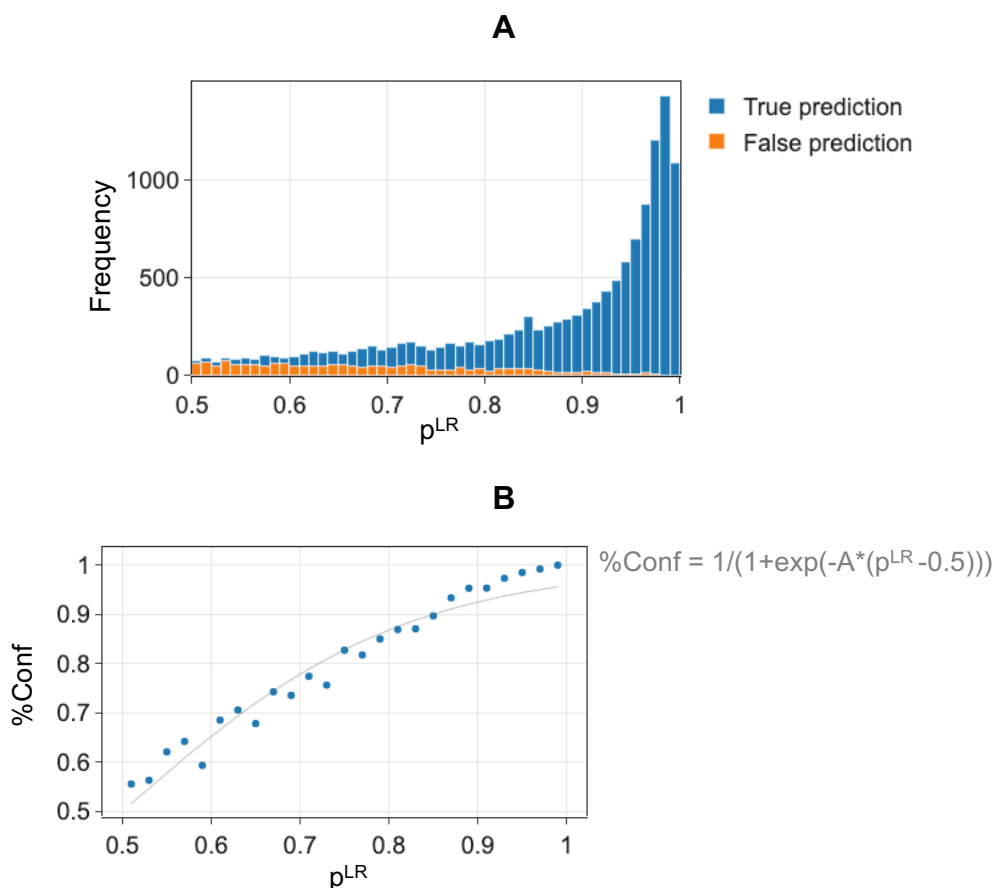


Figure 4.24. Derivation of the $\%Conf(p^{LR})$ function. A) Histograms of p^{LR} for true and false predictions; B) $\%Conf$ vs p^{LR} plot, $A = 6.29$

When given the conformational searching results at the FF level, a chosen priority list generation setting and outputs of partial re-optimisations at the DFT level, CONFPASS generates the χ_{new} vs r_{opt} plot and suggests whether further re-optimisation is needed with a confidence value, $\%Conf$. All the false predictions above $r_{opt} = 0.45$ are attributed to instances where the model predicts ‘stop’ when it should not. Thus, $1-\%Conf$ when $r_{opt} > 0.45$ is also equivalent to the percentage of false stop predictions.

$\%Conf$ values were calculated from the p^{LR} of the global LR model predictions from the cross-validation test. In Figure 4.25, the mean $\%Conf$ values from the test are plotted against r_{opt} for all 150 molecules. $\%Conf = 50\%$ is the margin between stop and continue predictions. On average, 14% ($r_{opt} = 0.135$) of the conformers need to be optimised to achieve $\%Conf = 50\%$. When $r_{opt} = 0.5$ (50% optimised), the average $\%Conf$ is 91%. For the most flexible molecules in the dataset (*ie* molecules with >170 conformers), $\%Conf$ at $r_{opt} = 0.5$ is over 95% (Figure 4.26). On average, $\%Conf$ no longer deviates below 90% after re-optimising 47% of the conformers. When $\%Conf > 90\%$, conformers contributing to approximately 86% of the Boltzmann distribution population should be covered (Appendix Table S4.3).

DFT dataset comprises distinct molecules that vary in terms of structure and flexibility. Specifically, for certain molecules like **4.13**, a %Conf score of 90% is achieved by re-optimising about 25% of the priority list. Conversely, for **4.14**, 70% of the conformers need to be re-optimised to exceed 90% %Conf.

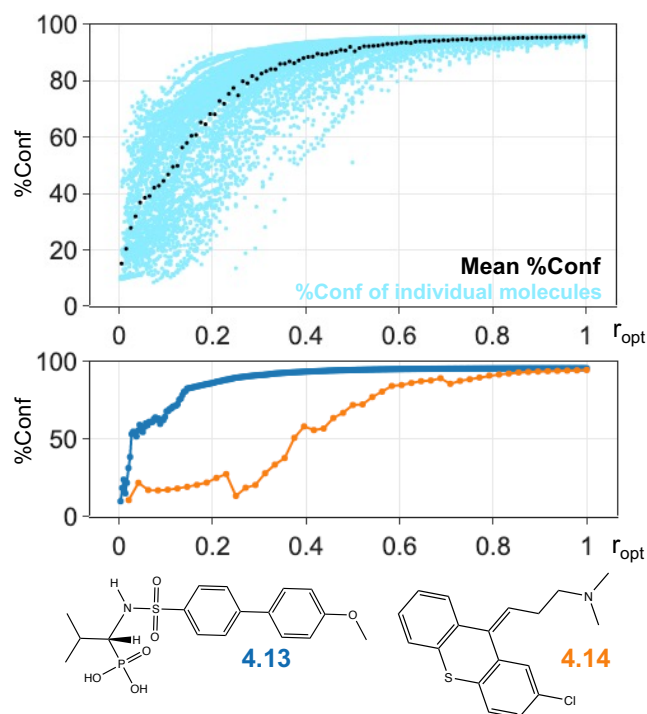


Figure 4.25. The percentage confidence (%Conf) vs r_{opt} plot: The light blue data points come from individual molecules in the DFT dataset. These data points were separated into 100 bins of equal size according to the r_{opt} values. The mean %Conf was found for each bin and plotted above in black. Examples of the %Conf vs r_{opt} plot for individual molecules are also presented.

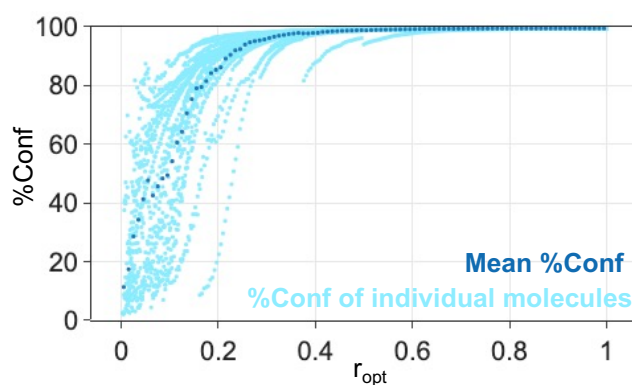


Figure 4.26. The combined %Conf vs r_{opt} plot and the mean %Conf vs r_{opt} plot for tests with priority lists from the *pipeline-mix* approach ($x = 0.8$, $Q = 0.2$) for the flexible molecules with more than 170 conformers. 24/150 molecules in the DFT dataset have more than 170 conformers.

4.6 CONFPASS at work

Based on the above studies, the CONFPASS program was developed to analyse conformational searching calculation outputs and guide the DFT re-optimisations. CONFPASS is fully automated, written entirely in Python and applicable for organic molecules, including radicals. The following packages were used: the Python (3.8.12) Standard Library, Pandas (1.3.4), Numpy (1.21.2), Sklearn (1.0.1),¹³² RDkit (2019.09.3)⁴² and Natsort (8.0.2). The work from Jensen *et al.* was also used to convert xyz coordinates to a RDkit mol object.^{152,153}

The CONFPASS program generates a priority list along with the input files to DFT calculations from the conformational searching result in SDF format. After partial re-optimisations of the FF structures at the DFT level, CONFPASS use a machine learning model to predict if the global minimum structure has been obtained with a percentage confidence value.

5 is a molecule collected by Hutchison *et al.* but outside the DFT dataset. Below, **15** is used as an example to demonstrate the application of CONFPASS in exploring conformational space (Figure 4.27, 4.28 and Table 4.2).

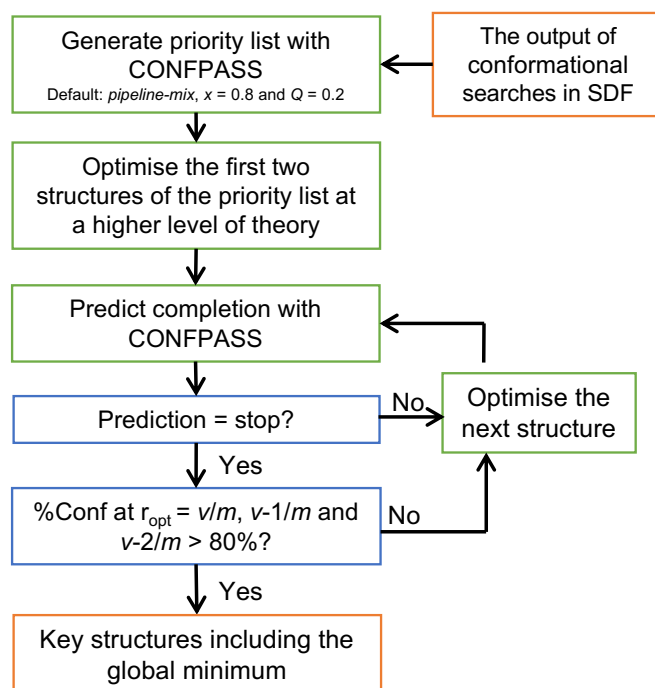


Figure 4.27. Usage recommendations of CONFPASS. m is the total number of conformers and v is the number of re-optimised conformers. The recommended cut-off for terminating the re-optimisation process should be higher than 80%. Depends on the computer system, the user of the CONFPASS may also wish to re-optimize FF structures in batches.

Table 4.2. CONFPASS at work: applications of CONFPASS in exploring the conformational space of **15** from the Hutchison dataset. Below is the complete version of the data table presented in the main text Figure 16. v is the number of re-optimised conformers. **15** has 129 conformers.

v	r_{opt}	Prediction	P^{LR}	%Conf	χ_{new}
2	0.016	Continue	0.52	47.5	0.000
8	0.062	Continue	0.60	34.2	0.268
15	0.116	Continue	0.60	34.5	0.073
21	0.163	Continue	0.58	37.6	1.000
28	0.217	Stop	0.55	57.9	0.000
34	0.264	Stop	0.51	51.4	1.000
41	0.318	Stop	0.64	70.7	0.000
47	0.364	Stop	0.72	79.6	0.000
52	0.403	Stop	0.77	83.8	0.000
53	0.411	Stop	0.78	85.3	0.000
54	0.419	Stop	0.79	85.8	0.000

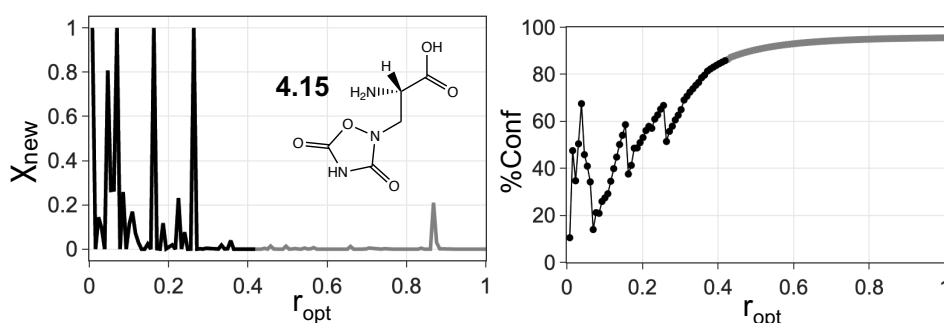


Figure 4.28. CONFPASS at work: applications of CONFPASS in exploring the conformational space of **4.15** from the Hutchison dataset. The black data points on the plots are known to the user at their wish after re-optimising 40% of the conformers following the default CONFPASS priority list (*pipeline-mix*, $x = 0.8$ and $Q = 0.2$).

Conformational searches of **4.15** give 128 conformers. A priority list is generated with CONFPASS using the default setting (*pipeline-mix*, $x = 0.8$ and $Q = 0.2$) and re-optimised the first two structures of the conformers following the priority list. At $r_{\text{opt}} = 2/129$, the LR model predicts further calculations are required. On the computer system in the Goodman group, it is more convenient to re-optimize conformers in batches. Thus, the next 5% of the conformers from the priority list were sent for DFT re-optimisations. After incorporating the updated results, a new prediction is generated using the LR model. The above procedure is iterated until the %Conf of the newly generated prediction surpasses 80%. At $r_{\text{opt}} = 0.42$ (ie 54/129), the %Conf exceeds 80%, indicating that the prediction of achieving the global minimal structure is highly probable. %Conf is also determined for $r_{\text{opt}} = 52/129$ and $r_{\text{opt}} = 53/129$ using P^{LR} from the prediction. Both values are greater than 80%, suggesting that a reliable and consistent

outcome has been achieved. The time required for DFT calculations can be reduced by running them in parallel or using cloud resources. This reduces the wall-clock time but not the computational resource requirement. CONFPASS make it possible to use the available computational resources more effectively.

A

```
## get the name of the sdf file in the same folder
path = os.getcwd()
sdf_ls = [f for f in os.listdir(path) if f.endswith('.sdf')]

## perform the clustering and generate the priority list
test1 = confpass.conp(sdf_ls)
test1.get_priority()

test1.priority_df
```

	idx	name	priority_ls
0	0	test_15.sdf	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 19, 16, 23,...
1	1	omegacsd_VORJER.sdf	[0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...

B

```
## Predicting the completion of the re-optimisation process
path = os.getcwd()+'/test_15'

test2 = confpass.pas(path)
test2.preparation()
test2.make_prediction()

reoptimisation: complete; confidence level: 83.08
probability: 0.753
current ropt (number of optimised / total number of conformers): 0.296
```

Figure 4.29. Using CONFPASS with Jupyter Notebook. A) Generating priority lists; B) Predicting the completion of the re-optimisation process

4.7 Comparison with CREST

CONFPASS is designed for obtaining all low-energy structures at a higher level of theory with a confidence level that all relevant structures have been found. CREST^{46,47} is often used for conformational studies and has a different aim: to generate representative structures, which is very useful for predicting spectroscopically relevant structures and finding macroscopic properties.

Conformational searching calculations were performed using CREST on 70 molecules with the default setting (Appendix to Chapter 4, 4.G.1). In the 70 molecules, 51 were randomly chosen from the Hutchison¹⁴⁷ dataset and 19 comes from the work of Grayson *et al.*⁵² All the conformers from the output files were re-optimised at the DFT level for a random selection of 20 molecules (*ie* the CREST-DFT dataset). 11 of the 20 CREST-DFT dataset molecules overlap with the DFT dataset (Figure 4.31), which allows us to compare the performance between CREST and MacroModel as the starting point of DFT re-optimisation processes.

DFT structures re-optimised from the CREST and MacroModel output for the 11 molecules were compared. In most cases (*ie* 9/11 cases), the same global minimum and low energy structures were identified upon DFT re-optimisations from the CREST and MacroModel output. However, conformational searches with MacroModel were more than 8 times faster compared to CREST (Figure 4.31.A). Although more conformer structures were found with CREST compared to MacroModel on average, 53% of the DFT re-optimised structures in the CREST-DFT dataset were duplicate structures. This means that CREST is much slower overall because it has to do many more DFT optimisations to get to the same outcome. Re-optimising all conformers from CREST searches needs 1.6 times more CPU time than re-optimising all the conformers from MacroModel. Using CONFPASS to give a greater than 90% confidence, with FF structures from MacroModel as the starting point, is more than 3.8 times faster than re-optimising all conformers from CREST outputs (Figure 4.31.B, see Appendix Table S4.6 and S4.7 for performance on the individual molecules).

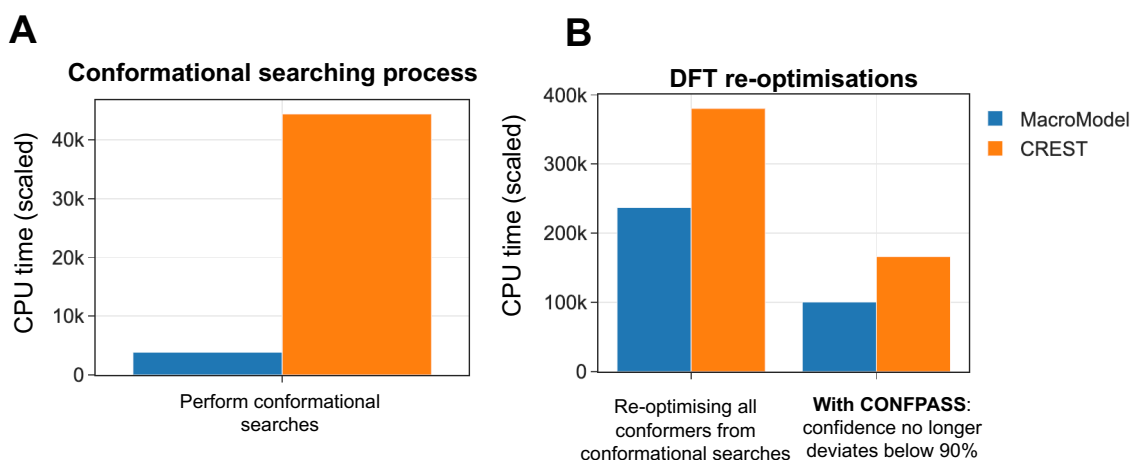


Figure 4.30. Performance comparisons by scaled CPU time. The values presented in above are the total time taken for the 11 molecules. A) The total CPU time for performing conformational searches is

compared for MacroModel and CREST. B) DFT re-optimisations were performed on the conformers from conformational searches. This bar chart compares the CPU time taken for re-optimising all conformers from conformational searches and performing the re-optimize with CONFPASS for MacroModel and CREST. For calculations with CONFPASS, the re-optimisation process is considered as completed when the confidence no longer drops below 90%.

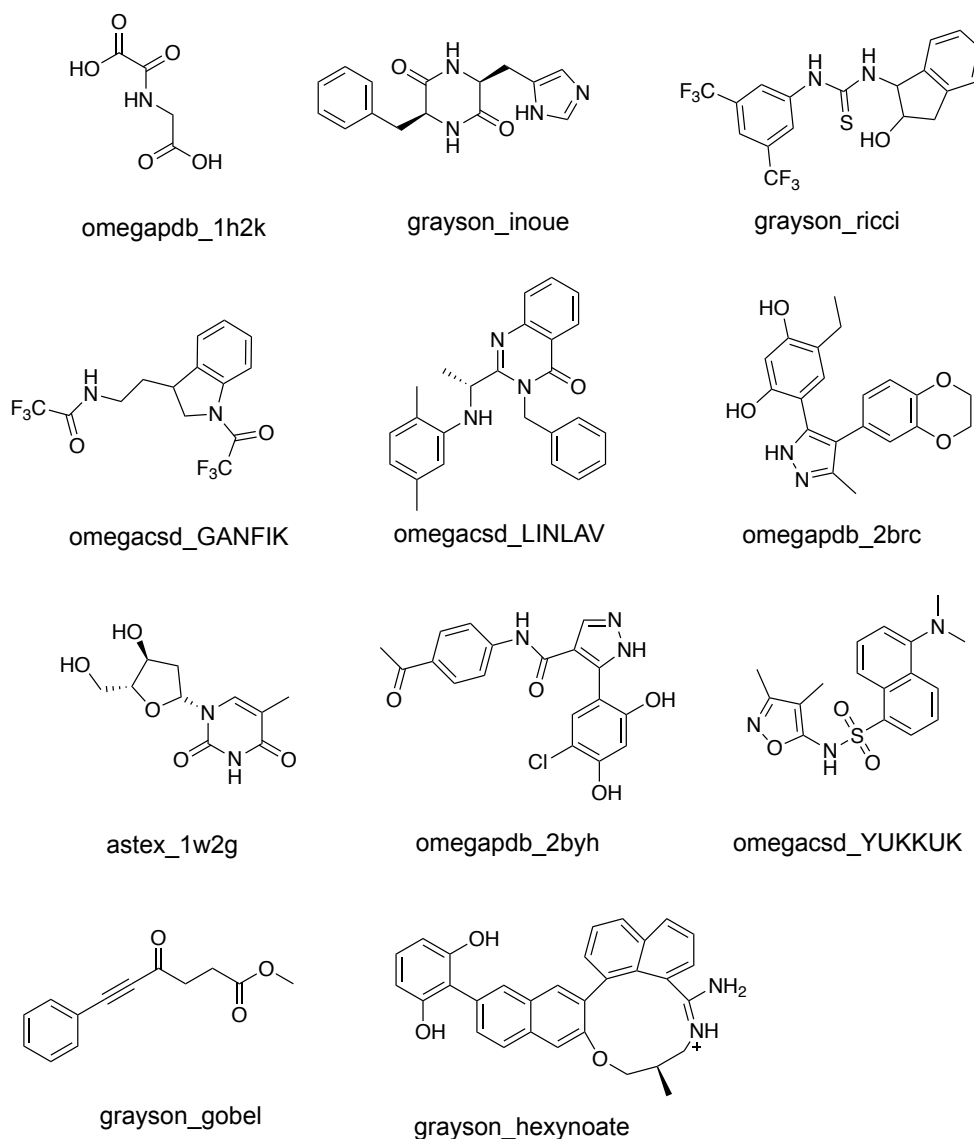


Figure 4.31. Performance comparisons between CREST and MacroModel: the structure of the 11 testing molecules

CONFPASS is flexible and able to accept conformational searching output files from any program that can generate SDF. Thus, CONFPASS can also accept output files from CREST. With the CREST-DFT dataset, the global minimum and bins test were performed and overall parameters were calculated on various priority list generation methods (Figure 4.32). Similar results were obtained compared to the result from the DFT dataset. Tests with the LR model were carried out for predicting the completion of the re-optimisation process with the CREST-DFT dataset. On average, CONFPASS gives more than 90% confidence after re-optimising 43% of the conformers following the default priority list, which covers 90% of the Boltzmann distribution population (Table S4.8). The above investigation demonstrates the robustness of CONFPASS in accepting output files from calculations at different levels of theory and conformational searching methods. The prioritisation and pas test prediction from CONFPASS can help obtain the global minimum and key low-energy structures with a reduced computational cost.

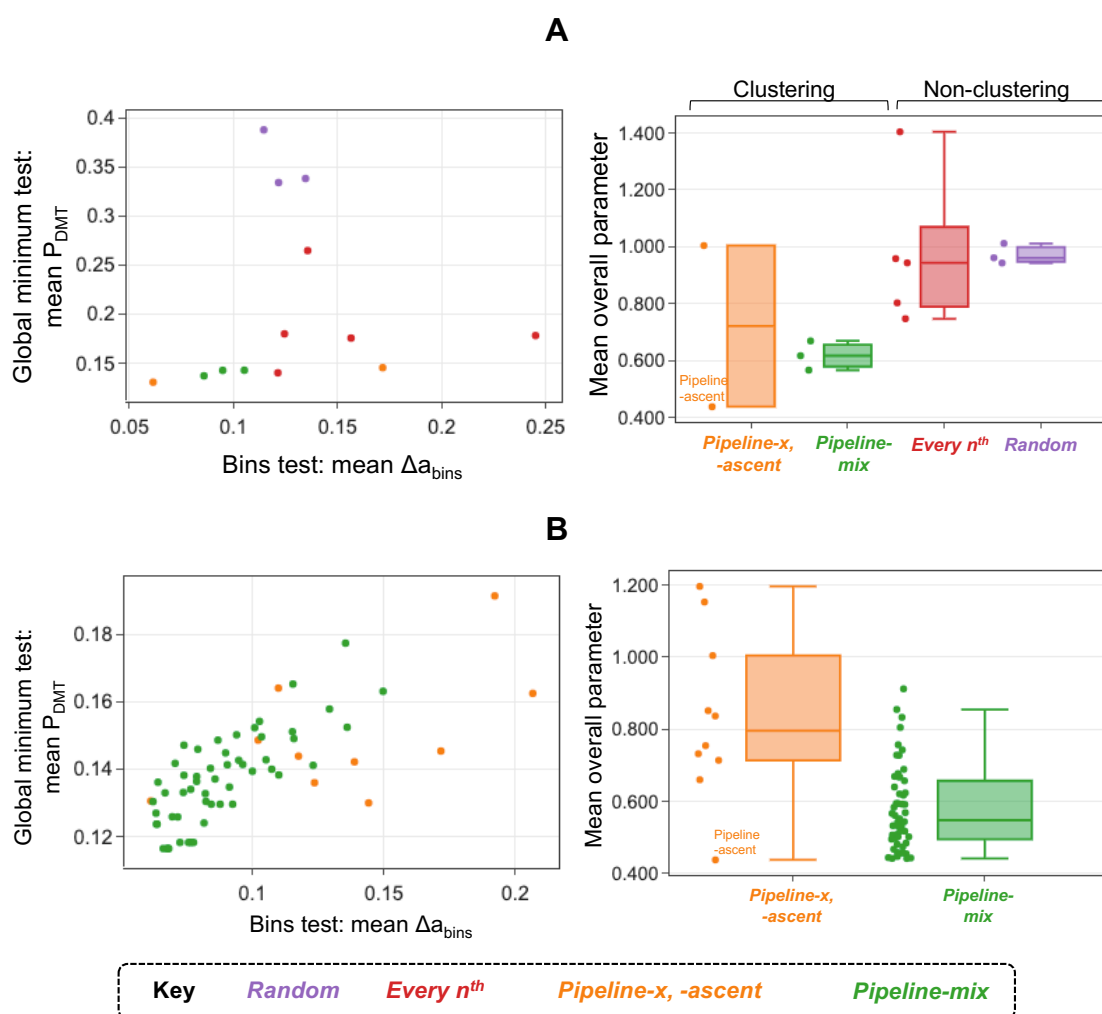


Figure 4.32. The global minimum and bins test results with the CREST-DFT dataset. A) Performance of the clustering approach pipeline with CREST conformational searching outputs as the starting point. The tests for the *random* approach have been repeated for 5 times, which contributes to the 5 data points under this category. B) Hyperparameter tuning on the x and Q parameter for the *pipeline-x* and *pipeline-mix* approach: The global minimum test (P_{GMT}) and bins test (Δa_{bins}) results are presented in the scatter plot. The mean overall parameter data is presented as the box plot. The data points on the

scatter plots and box plots are coloured according to the corresponding priority list generation approach. Hyperparameter tuning tests were conducted on the x and Q parameter for the *pipeline-x* and *pipeline-mix* approach. The *pipeline-mix* approach with $x = 0.2$ and $Q = 0.15$ gives the best (lowest) overall parameter value although this is only marginally better than the *pipeline-ascent* approach.

If the aim is to find representative DFT geometries, CREST is the appropriate choice. If you need the lowest energy DFT structures and a confidence level that you have found them, then CONFPASS gives this information faster and more effectively.

4.8 Summary

In conclusion, a practical framework for analysing conformational searching output and generating priority lists for re-optimising conformers at the DFT level has been established. The automated clustering approach pipeline extracts dihedral angle values as descriptors from conformers, performs the agglomerative clustering process and generates the priority list. Thorough evaluations were conducted through the global minimum test and the bins test based on a large dataset of 150 structurally diverse molecules. The priority list from the pipeline intuitively guides the DFT re-optimisation process and reduces the chance of getting duplicate conformers by up to 51% for the first 30% of the calculations. Machine learning models for predicting the completion of the re-optimisation process were trained and evaluated. Users can finish their DFT re-optimisation process at an earlier stage with the confidence that the global minimum has been obtained based on the p^{LR} value.

This project led to the development of the CONFPASS program, which covers the clustering approach pipeline for generating priority lists and tests for predicting the completion of the re-optimisation process. CONFPASS is fully automated and offers rapid analyses. CONFPASS has been tested with FF structures from MacroModel and semi-empirical structures from CREST, which demonstrates the robustness of the program.

The CONFPASS program is available to download on GitHub (github.com/Goodman-lab/CONFPASS), with a detailed user guide and examples. The scripts can be executed as a single command line or imported and used in Python scripts. The analyses are rapid and take on average few seconds to process an SDF or a Gaussian 16 output file folder.

In the future, the CONFPASS program will be further developed to cope with organometallics and multimeric systems. Adapting the program for finding transition states is also an interesting aspect to investigate.

Chapter 5

Oxazaborolidinium ion-catalysed reactions: Reaction dynamics as the missing puzzle piece

5.1 Background

Sometimes, thorough explorations of the possible transition state geometries are not enough to predict the selectivity of an organic reaction. This chapter highlights the practical relevance of dynamic effects to experimental organic chemistry. VRAI-selectivity, recently developed within the Goodman group, can analyse and predict selectivity outcomes of dynamically controlled reactions without expensive computations.

In this chapter, VRAI-selectivity calculations were performed to investigate a set of complex organic reactions reported by Ryu *et al.*^{154–156} and demonstrate that their selectivity is heavily influenced by reaction dynamics (Figure 5.1.A). Notably, the chemical systems in this investigation feature significant complexity, encompassing a range of 90–112 atoms. The reactions are catalysed by chiral oxazaborolidinium ion (COBI) catalysts, a well-established and versatile catalyst system developed by Corey *et al.*^{157,158} COBI catalysts effectively activate carbonyl group as a Lewis acid and thus have broad applicability in various organic reactions under research setting.^{159–163} The study initially centres around the ketone-selective reactions (Figure 5.1.B).¹⁵⁶ These reactions involve identical substrates but exhibit distinct chemo- and stereo-selective outcomes due to subtle variations in the substituents on the aromatic rings of the catalysts. The result shows that reaction dynamics plays a key role in determining the selectivity outcome. To validate this finding, two additional reactions were further selected: the epoxide¹⁵⁴- and aldehyde¹⁵⁵-selective reaction (Figure 5.1.C and D), in which the main products are epoxide and aldehyde, respectively. Despite changes in the substituents of diazo and aldehyde substrates, the underlying mechanism remains the same and the reaction outcomes are also governed by reaction dynamics.

The publications of previous computational work in understanding the reactivity of COBI catalysts are noted.^{164–171} COBI-catalysed Diels-Alder reactions have been studied extensively by several research group.^{167–169} These studies have focussed on rationalising the selectivity based on structural features of the key transition state (TS) structures and non-covalent interactions between the substrates.^{164,165} To the best of my knowledge, the selected reactions have not been studied considering both reaction dynamics and the full conformational space of the system.

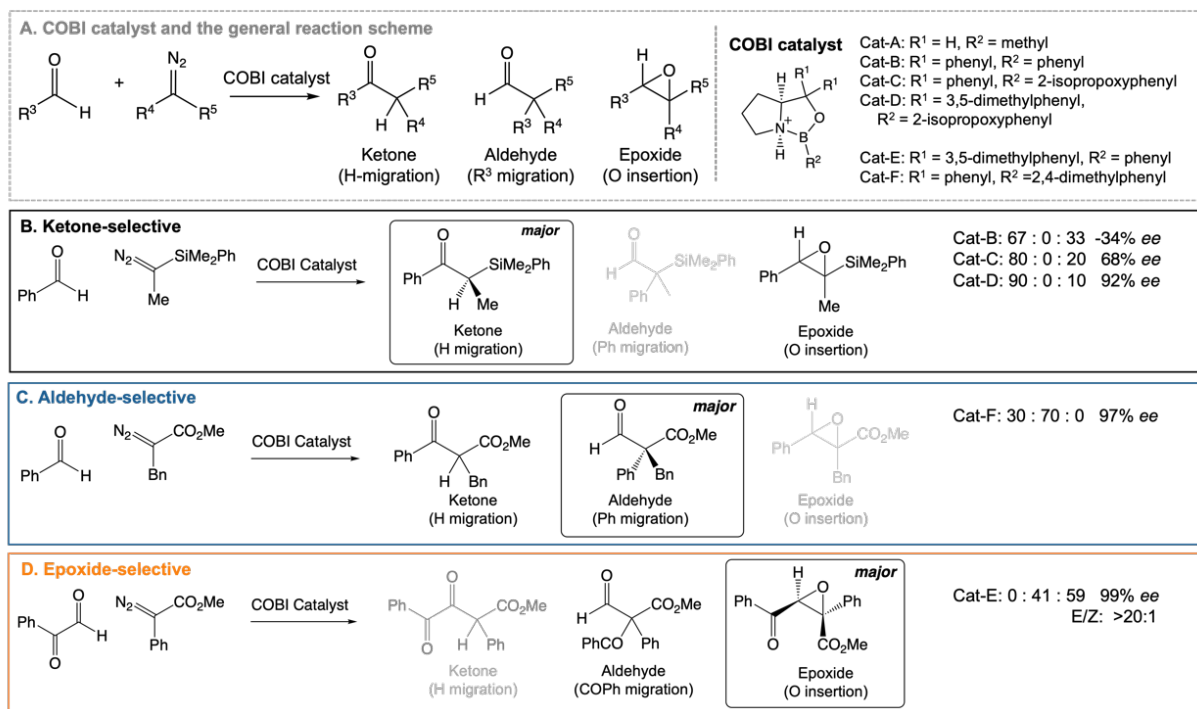


Figure 5.1. Chiral oxazaborolidinium ion (COBI) catalysed reactions between aldehyde and diazo compounds.^{154–156} The above reaction systems were investigated in this project. All enantiomeric excess (ee) values above are determined from chiral HPLC. For the ketone¹⁵⁶- and aldehyde¹⁵⁵-selective reaction, the major to minor product ratio comes directly from the crude product ¹H NMR. For the epoxide-selective reaction,¹⁵⁴ the yield for the aldehyde has not been explicitly reported for this reaction. It is assumed that the loss in percentage yield of the epoxide is all attributed to the aldehyde by-product. The major to minor product ratio is estimated from percentage yield of the isolated *SR* epoxide and diastereomeric ratio, which comes from ¹H NMR of the crude product. Ryu *et al.* reported that changing the diazo substituent from CO₂^tBu to CO₂Me do not alter the experimental outcome significantly. Thus, the diazo substituent from CO₂^tBu was simplified to CO₂Me for the computational investigation.

The selectivity controlled by reaction dynamics in complex organic systems is an underexplored area. The findings from this study showcase the efficacy of VRAI-selectivity in filling this gap in our understanding of selectivity and provide more accurate insights into the reaction process and help to open up new avenues for developing better synthesis strategies.

5.2 Computational methods

Conformational searching calculations were conducted in MacroModel (V13.4)³⁹ with OPLS4¹⁶ force field and a maximum of 2000 steps as the limit. DFT calculations were conducted with Gaussian 16.¹⁰⁴ All optimisations were performed at the B3LYP-D3/6-31G(d) level of theory.^{105–108} Single-point energy calculations were carried out at the ω B97X-D/6-311G(d,p) level of theory.¹⁰⁹ Conformers with energies within a window of 41 kJ mol⁻¹ (equivalent to 10 kcal mol⁻¹) were saved for subsequent analyses.

The conformational space of key ground state structures was explored thoroughly. CONFPASS was used to assist the DFT re-optimisations of force field structures with confidence that key stable structures are obtained at a reduced computational cost (Chapter 4).^{172,173} The default setting (*pipeline-mix*, $x = 0.8$ and $Q = 0.2$) was used for generating the priority list for re-optimising conformers. It was ensured that the %Conf at r_{opt} is greater than 80% before terminating the re-optimisation process. On average, there are more than 87% confident that the global minimum structure has been obtained after re-optimising less than 37% of the conformers from the conformational searching output file (Figure 5.2).

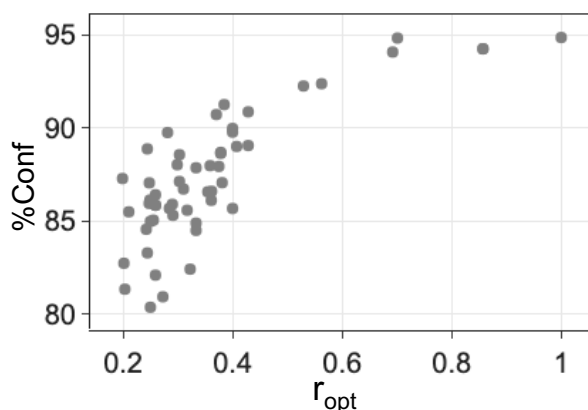


Figure 5.2. Distribution of the final %Conf and r_{opt} from the CONFPASS test for confirming the completion of the re-optimisation process (*ie* the global minimum has been obtained).

The VRAI-selectivity algorithm⁶⁴ was utilised to account for the effects of reaction dynamics on reaction selectivity. Before running the main algorithm, an energy check was run at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) level of theory to ensure that the second transition states (TS^{II}s) are lower in Gibbs free energy than the first transition state (TS^I). Otherwise, the programme would proceed to calculate the product percentages with the transition state theory (TST). The main VRAI-selectivity algorithm takes the geometries and frequency of the TS^I, intermediate and products as inputs. The frequency analyses were conducted at the B3LYP-D3/6-31G(d) level of theory in this study. The dimensionality of the potential energy surface (PES) is reduced to two dimensions by examining the bond differences between the products or between a product and TS^I based on their geometries (Figure 5.3). The major product is determined by the direction in which the imaginary eigenvector of TS^I points relative to the intermediate on the 2D PES. The width of the trajectory stream is estimated based on

the real eigenvectors of a TS using harmonic oscillator approximation. The selective ratio is predicted by considering the width of the trajectory stream at TS^I and how much of it favours each product.

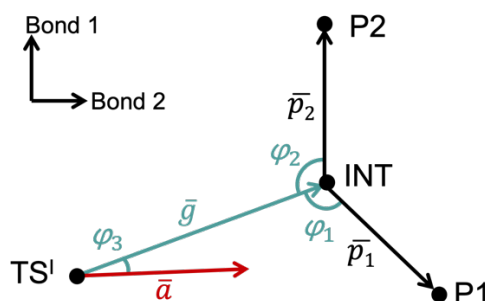


Figure 5.3. A qualitative illustration of the 2D PES projection from the VRAI-selectivity calculation. $\bar{a}$ is the imaginary eigenvector of TS^I. $\bar{g}$ is the separation between TS^I and INT. $\bar{p}_1$ and $\bar{p}_2$ are the displacement vectors from INT to product 1 (P1) and 2 (P2), respectively.⁶⁴ φ_1 and φ_2 are the angle that $\bar{g}$ makes with $\bar{p}_1$ and $\bar{p}_2$ respectively. φ_3 is the angle between $\bar{g}$ and $\bar{a}$.

In this study, an extension to VRAI-selectivity, VRAI-multi, was developed and applied. VRAI-selectivity is designed to be executed on terminals and for dealing with only one set of input files with two products at a time. VRAI-multi automates the process of VRAI-selectivity analyses in situations where there are more than two different products sharing the same intermediate and TS^I on their reaction pathways. VRAI-multi is written as a Python package, which allows more flexibility in tuning hyperparameters, such as temperature, and integrations into other Python programmes. The scripts are available at github.com/Goodman-lab/VRAI-selectivity.

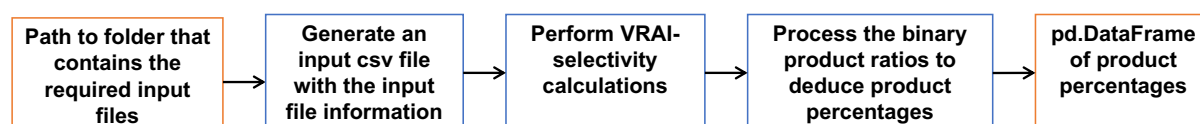


Figure 5.4. The process flow chart for VRAI-multi.py

The input to VRAI-multi is the path to the folder that contains the required files for the VRAI-selectivity analyses (elaborated in Appendix to Chapter 5, 5.A). VRAI-multi generates a data frame that comprises rows containing input file information for conducting VRAI-selectivity calculations. The data frame is constructed to encompass all possible binary combinations of products and TS^Is.

Following each row in the input file data frame, the VRAI-selectivity calculations are executed. Binary product ratios are obtained and recorded into a raw result data frame. The raw result table is processed to calculate the product percentages. For results that involve the same intermediate and TS^I:

1. The calculation results that share a common product are identified.

2. The product percentages are determined by calculating the ratio of the common product in different sets of results.
3. The above process is repeated for all possible products.

The output of the pipeline is a data frame that contains the product percentages, which are calculated by using different products as the reference in the calculation.

5.3 The mechanism

Firstly, the uncatalysed pathways were studied using substrates from the ketone-selective reaction. The substrates were simplified by replacing phenyl with methyl groups. The mechanism for the uncatalysed pathway is distinctively different from the COBI-catalysed pathway. The C-C formation leads to cyclic intermediate structures. The pathway via the 5-membered ring intermediate is more favourable kinetically compared to the pathway via the 4-membered ring intermediate.

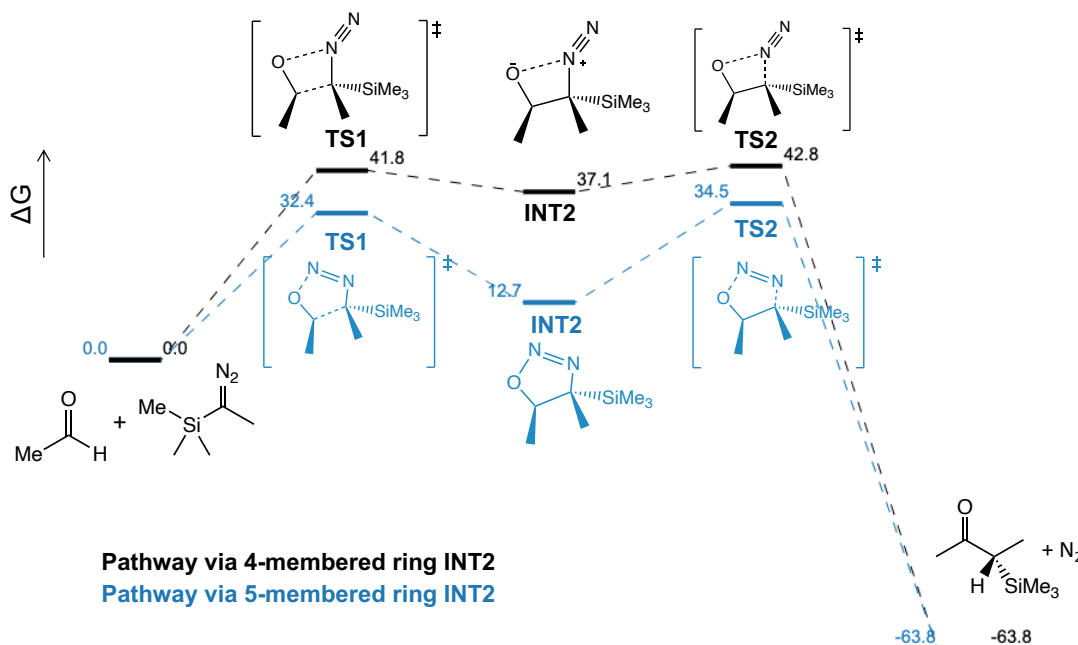


Figure 5.5. The energy profile of the uncatalysed pathway. The ΔG values relative to the reactant are labelled on the diagram in kcal mol^{-1} . The stereochemistry in INT2 is *RR*.

For the catalysed pathway, Wei *et al.* have previously conducted a mechanistic study on the chosen aldehyde-selective reaction (Figure 5.1.C).¹⁷¹ A preliminary mechanistic study was performed with a simplified COBI catalyst (*ie* Cat-A, Figure 5.1). The results are applicable to all other reactions and these conclusions align with those of Wei *et al.* (Figure 5.6) The complexation between the COBI catalyst and the aldehyde leads to INT1, where the formed B-O bond is *syn* to the N-H bond. The ΔG of the complexes with the opposite stereochemistry at the boron atom are at least $3.3 \text{ kcal mol}^{-1}$ higher compared to INT1 across the various reaction systems in this study (Appendix Figure S5.2). The addition of the diazo on INT1 gives an adduct, INT2, which undergoes a nitrogen elimination. Sequentially, a 1,2 shift or oxygen insertion occurs to give the products, which can be a ketone, aldehyde or epoxide.

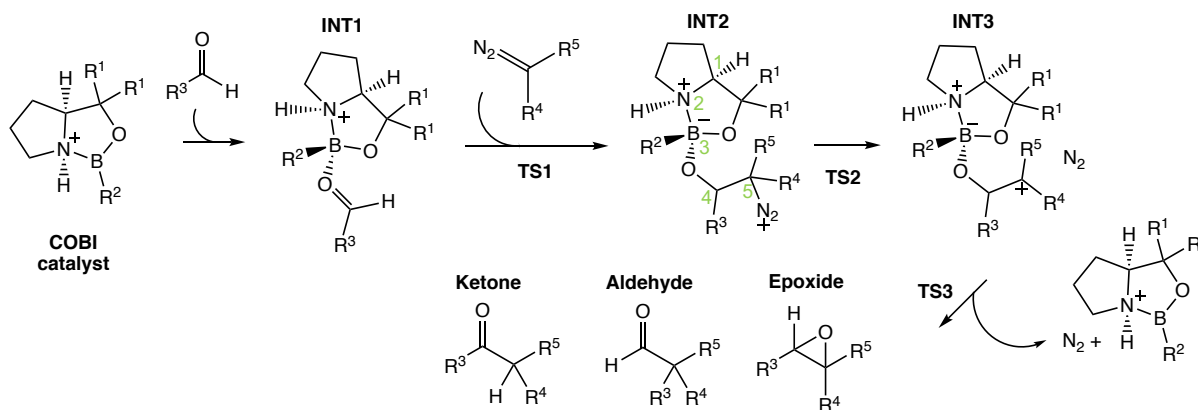


Figure 5.6. General mechanism for the COBI-catalysed reactions. The key stereo centres in INT2 are numbered.

The carbocation structure after the nitrogen elimination, INT3, presents as a shoulder on the PES and the sequential process tends to be barrierless (*ie* $\Delta G^\ddagger(\text{TS3}) = 0 \text{ kcal mol}^{-1}$). The group involved in the 1,2 shift or oxygen insertion must be at an anti-periplanar position relative to the leaving nitrogen. Hence, the stereochemistry of the product is determined by the stereochemistry of INT2 at position C4 and C5. The chemoselectivity is controlled by the process after INT2. Thus, the focus of this investigation for understanding selectivity will be on the addition between diazo and INT1 via TS1 and nitrogen elimination via TS2.

There are some exceptions to the trend described above in the epoxide-selective reaction. The differences will be discussed in the following sections.

5.4 Predicting selectivity with VRAI-Selectivity

The product percentage was predicted for the reaction systems presented in Figure 5.1 based on calculations and comparisons to the experimental results.

Here, the ketone-selective reaction with Cat-B will be used as an example to elaborate on the procedure for calculating the product percentage. The transition state theory (TST) was considered. It was assumed the product percentages are controlled kinetically by TS1s and TS2s. The process of the calculation is illustrated with the Sankey diagram in Figure 5.7.B. The percentage populations via the TSs were computed from the $\Delta\Delta G^\ddagger$. Only the low-energy TSs, whose combined population accounted for 99% of the Boltzmann distribution, were selected for the calculation. The final product percentage was determined by summing up the percentage populations of the pathways that contribute to each product. Distinct mismatches between the experimental and calculated product percentages were found (Table 5.1). TST predicts the aldehyde as the major product and contributes to more than 50% of the final product compositions.

The large difference between the TST predictions and the experimentally observed selectivity has been noted. One possible explanation for this is that the reaction process is partially controlled by reaction dynamics. However, it is also possible that important transition states have been missed out. To check the second possibility, the conformational space was explored thoroughly for INT2 using CONFPASS. All possible conformations of TS1 and TS2 were optimised based on the INT2 structures at the DFT level for the ketone-selective reaction with Cat-B. Hence, mismatches with the experimental results were not due to insufficient conformation sampling. Secondly, the energy profile (Figure 5.7.A) reveals features of PES where dynamic effects dominate and control the selectivity. The drop from TS1 to INT2 is more than 8 kcal mol⁻¹, while the activation energy for TS2 is less than 3 kcal mol⁻¹. This suggests that there is sufficient energy for trajectories to pass over INT2 easily and the process beyond TS1 may be controlled by reaction dynamics. This assumption is cooperated into the analyses and tested by comparison with the experimentally determined outcomes.

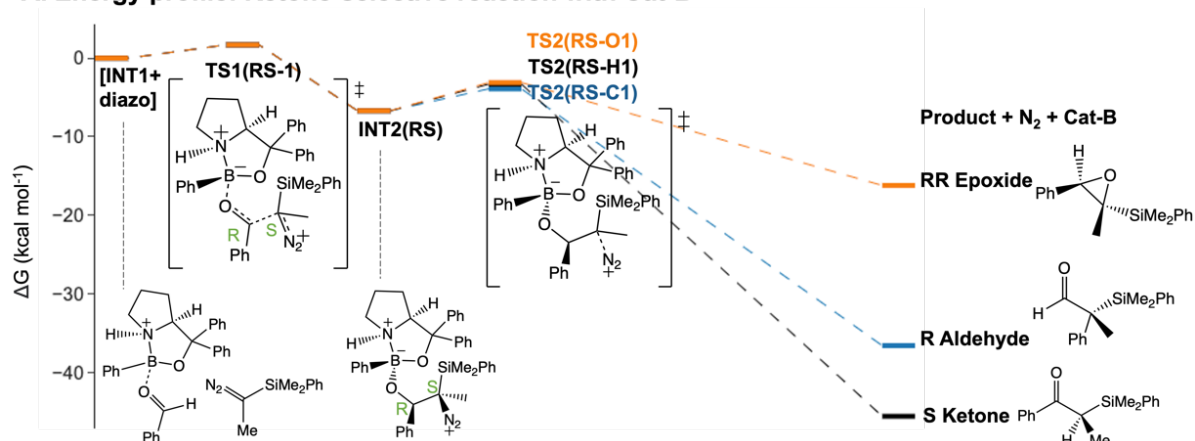
The effects of reaction dynamics were investigated using the VRAI-selectivity approach (Figure 5.7.C). A VRAI-selectivity calculation is performed for every selected low-energy TS1. The analyses were performed with TS1 as the first transition state (TS¹) and INT2 as the intermediate structure. The product geometries come from the quick reaction coordinate calculations¹⁴¹ of the lowest energy TS2 structure that leads to the aldehyde, ketone or epoxide. The percentage population for each pathway was obtained by multiplying the product percentage from VRAI-selectivity analyses and the percentage population of the TS1 from $\Delta\Delta G^\ddagger$. By accounting for the effects of reaction dynamics, the major product chemoselectively and stereoselectively has been predicted successfully for the reaction with Cat-B. The energy profiles for the ketone-selective reaction with Cat-C and Cat-D are similar to that of the reaction with Cat-B. The product percentage calculation with and without considering reaction dynamics was repeated. The predicted percentages incorporating VRAI-selectivity calculations show much better

agreement with the experimental results and suggest that the chemoselectivity is significantly influenced by reaction dynamics.

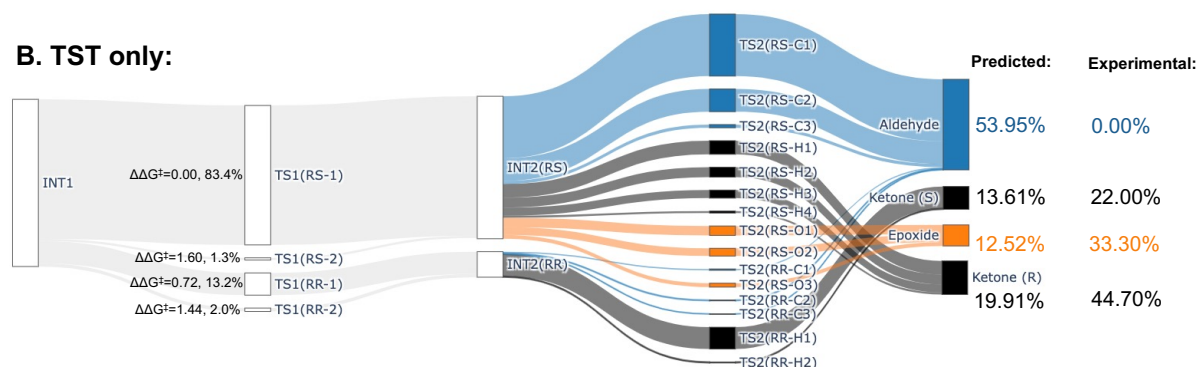
Table 5.1. Comparisons to the experimental results. ‘Calculated percentage (TST)’ is derived based on the TST assumption. The selectivity of the reaction entirely depends on the kinetics of TS1 and TS2. ‘Calculated percentage (with VRAI)’ incorporates the VRAI-selectivity calculation outcomes. It is assumed that the stereochemistry is controlled by kinetics via TS1 and the chemoselectivity, *ie* processes beyond TS1, are controlled by reaction dynamics.

	Experimental Percentage	Calculated Percentage (TST)	Calculated Percentage (with VRAI)
Ketone-selective – Cat-B			
Aldehyde	0.00%	53.95%	3.32%
Epoxide	33.30%	12.52%	4.33%
Ketone (<i>R</i>)	44.70%	19.91%	80.45%
Ketone (<i>S</i>)	22.00%	13.61%	11.90%
Ketone-selective – Cat-C			
Aldehyde	0.00%	31.16%	15.41%
Epoxide	25.00%	0.00%	21.41%
Ketone (<i>R</i>)	12.00%	0.00%	9.53%
Ketone (<i>S</i>)	63.00%	68.84%	53.65%
Ketone-selective – Cat-D			
Aldehyde	0.00%	14.34%	11.34%
Epoxide	11.11%	0.00%	6.73%
Ketone (<i>R</i>)	3.56%	0.00%	2.80%
Ketone (<i>S</i>)	85.33%	85.66%	73.08%
Epoxide-selective – Cat-E			
Aldehyde	40.91%	35.80%	44.03%
Epoxide (<i>SR</i>)	56.00%	11.90%	38.86%
Epoxide (<i>RS</i>)	0.28%	0.00%	0.00%
Epoxide (<i>SS + RR</i>)	2.81%	52.15%	3.87%
Ketone	0.00%	0.15%	10.32%
Aldehyde-selective – Cat-F			
Aldehyde (<i>S</i>)	1.04%	0.00%	0.00%
Aldehyde (<i>R</i>)	68.95%	8.64%	54.33%
Epoxide	0.00%	0.00%	0.00%
Ketone	30.00%	91.36%	45.67%

A. Energy profile: Ketone-selective reaction with Cat-B



B. TST only:



C. With VRAI-selectivity:

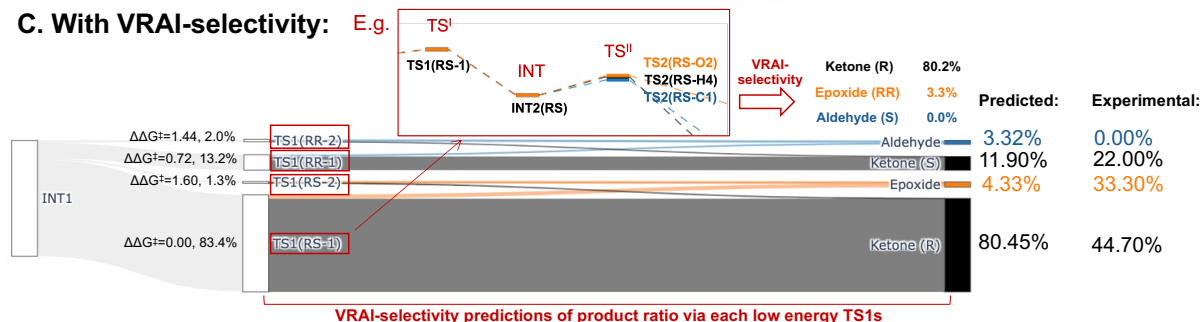
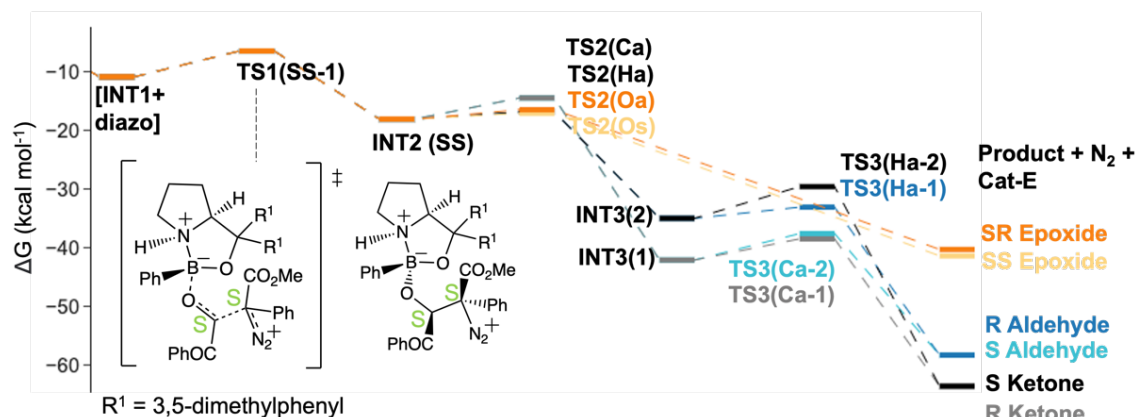


Figure 5.7. Product percentage calculations for the ketone-selective reaction with Cat-B. A) The energy profile of the major pathways via TS1(RS-1), which contribute to 83.4% of the products in the final composition, is presented at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) level of theory. The stereochemistry of INT2 at C4 and C5 is RS. B) The Sankey diagram illustrating the percentage calculation of products, based on transition state theory (TST) only, is presented. C) The predicted product percentages show a better match with the experimental result after incorporating VRAI-selectivity calculations. A VRAI-selectivity calculation is conducted for every pathway via a low-energy TS1 (*ie* highlighted in red box). The structures on the diagram are named according to the C4 and C5 stereochemistry and numbered based on their energy ranking. For example, TS2(RS-C1) is the lowest energy RS TS2 that leads to an aldehyde product.

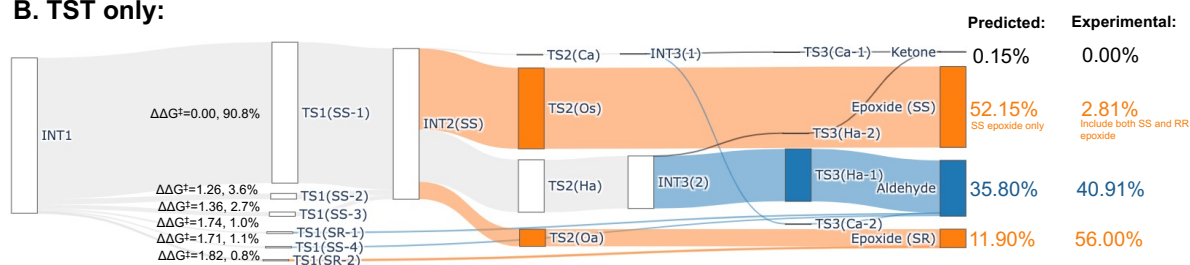
The PES of the epoxide-selective reaction is noticeable different from the ketone-selective reactions (Figure 5.8.A and B). Firstly, in minor pathways via TS1(SS-4), TS1(SR-1) and TS1(SR-2), the process beyond TS1 is barrierless and leads to the final product directly. Secondly, in the major pathway (*ie* via TS1(SS-1)), the *syn*-periplanar insertion (*ie* the inserting shifting group at the *syn*-periplanar position to the leaving nitrogen) has comparable kinetic barrier to the anti-periplanar insertion. The TS1 structure with the lowest $\Delta G^\ddagger$, TS1(SS-1), contributes to both SR and SS epoxide. Unlike in other reactions, INT3 structures presents as a stable intermediate in the energy profile ($\Delta G^\ddagger(\text{TS3}) > 0 \text{ kcal mol}^{-1}$) and can lead to both ketone (*R/S*) and aldehyde (*R/S*) upon 1,2-shift of the key group. Thus, all possible six products were considered in the VRAI-selectivity calculations for this reaction. The product percentages were calculated with and without incorporating VRAI-selectivity results. The predicted percentages with VRAI-selectivity calculations show a much closer match with the experimental result. The same conclusion as for the ketone-selective reaction can be drawn. The stereoselectivity is controlled kinetically by TS1, while the chemoselectivity is determined by the reaction dynamics of the process beyond TS1.

The selectivity of the aldehyde-selective reaction was found to be controlled by reaction dynamics entirely. The TS1(*RR*) with the lowest $\Delta G^\ddagger$, with a stereochemistry of *RR* at C4 and C5 position, do not lead to the major product, *R* aldehyde. The earlier study of Wei *et al.*, the selectivity is rationalised based on TST.¹⁷¹ Starting from the TS1 with the lowest $\Delta G^\ddagger$ from their publication, which has a *RS* stereochemistry at C4 and C5, the system was investigated using a functional with a dispersion correction and thorough conformational exploration. This led to a different conclusion from TST which was now incompatible with the experimental result (Table 5.2). Therefore, further investigations were conducted to determine whether the stereochemistry of the reaction may also be controlled by reaction dynamics. The drop in energy from TS0, the TS for the COBI catalyst and aldehyde complexation, to INT1 is more than 10 kcal mol⁻¹. TS0 is more than 2.5 kcal mol⁻¹ higher compared to the lowest energy TS1 (Figure 5.9.A). An additional VRAI-selectivity calculation was conducted with TS0 as the TS^I and INT1 as the intermediate. The products are stereoisomers of INT2, INT2(*RR*) and INT2(*RS*). As the lowest energy TS2 to the epoxide is higher in energy than the corresponding TS1, this pathway is considered to be unfavourable kinetically. The final product percentages with VRAI-selectivity results were calculated based on three sets of VRAI-selectivity calculations. The Sankey diagram in Figure 5.9.C provides the details to the calculation. The calculated percentages incorporating VRAI-selectivity calculation results show a better match with the experimental result than the calculated percentages based on TST (Table 5.1). This implies that both the stereochemistry and chemoselectivity in the aldehyde-selective reaction may be influenced by reaction dynamics.

A. Energy profile: Epoxide-selective reaction



B. TST only:



C. With VRAI-selectivity:

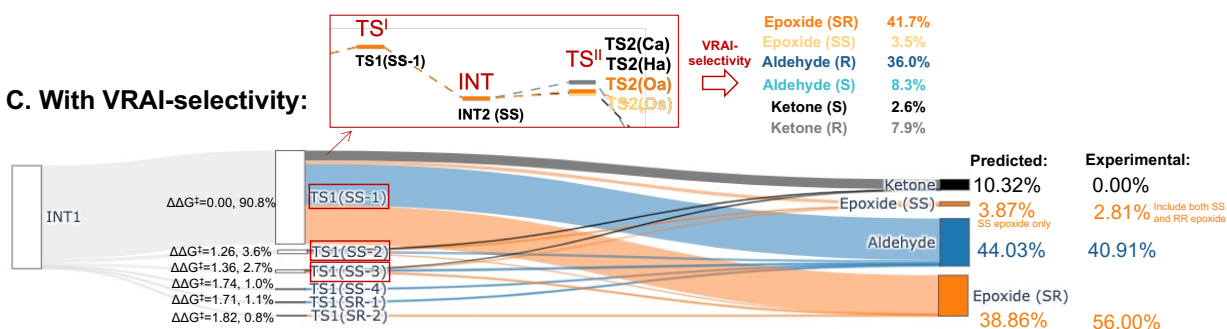
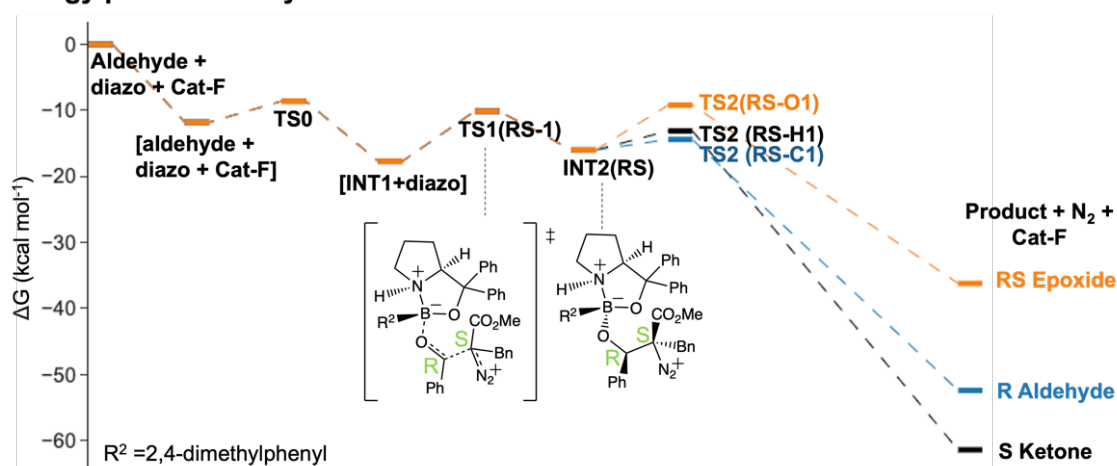
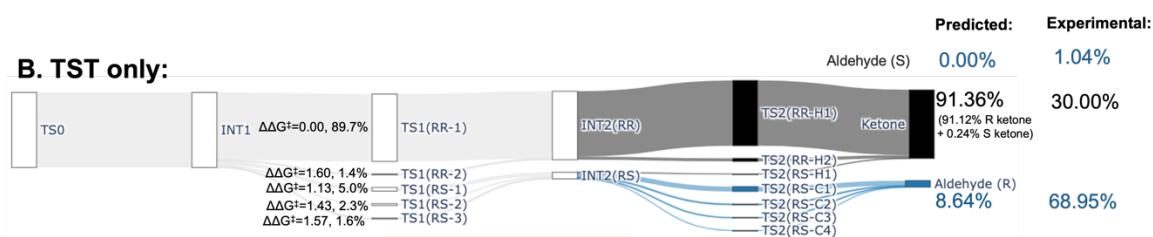


Figure 5.8. Product percentage calculations for the epoxide-selective reaction with Cat-E. A) The energy profile of the major pathways via TS1(SS-1) is presented at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) level of theory. The stereochemistry of INT2 at C4 and C5 is SS. B) The Sankey diagram illustrating the percentage calculation of products, based on transition state theory (TST) only, is presented. C) The predicted product percentages show a better match with the experimental result after incorporating VRAI-selectivity calculations. Three sets of VRAI-selectivity calculations were conducted with TS1(SS-1), TS1(SS-2) and TS1(SS-3) as the first transition state (*ie* highlighted in red box in the Sankey diagram). The structures on the diagram are named according to the C4 and C5 stereochemistry position and numbered based on their energy ranking. Experimentally, there are also 0.28% of the *RS* epoxide in the final composition, which is omitted in the above diagram. The calculations, with and without VRAI-selectivity, predict that no *RS* epoxide should be obtained.

A. Energy profile: Aldehyde-selective reaction



B. TST only:



C. With VRAI-selectivity:

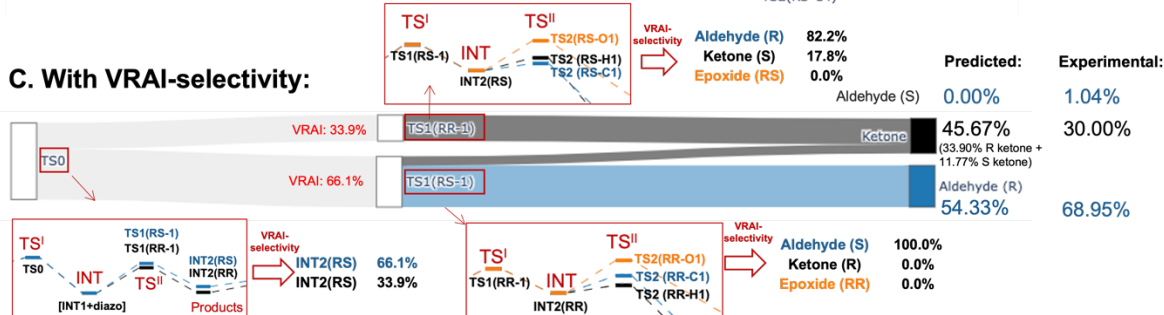


Figure 5.9. Product percentage calculations for the aldehyde-selective reaction with Cat-F. A) The energy profile of the major pathways via TS1(RS-1) is presented at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) level of theory. The stereochemistry of INT2 at C4 and C5 is RS. B) The Sankey diagram illustrating the percentage calculation of products, based on transition state theory (TST) only, is presented. C) The predicted product percentages show a better match with the experimental result after incorporating VRAI-selectivity calculations. Three sets of VRAI-selectivity calculations were conducted to determine the final predicted product percentages. Here, the stereoselectivity is also controlled by reaction dynamics, which differs from other reactions. The structures on the diagram are named according to the C4 and C5 stereochemistry and numbered based on their energy ranking.

Table 5.2. Comparisons with previous studies from Wei *et al.* Wei *et al.* have previously conducted a mechanistic study on the chosen reactions for the aldehyde-selective reaction at the B3LYP/6-31g(d)/propionitrile/PCM level of theory. The xyz coordinates of the key TS1 structures were extracted from the supporting information of their work. The structures were re-optimised and $\Delta G^\ddagger$ values were calculated with the theory level used in this study, *ie* ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d).

$\Delta\Delta G^\ddagger(\text{TS1}) / \text{kcal mol}^{-1}$				
Stereochemistry at C4 and C5	<i>RR</i>	<i>RS</i>	<i>SR</i>	<i>SS</i>
This work	0.00	1.12	4.54	4.80
Wei <i>et al.</i>	6.69	1.57	12.58	8.56

- R^1 = phenyl, R^2 = 2,4-dimethylphenyl

Inclusion of SMD solvent models^{130,174} in single-point energy calculations and structure re-optimisations with other functionals (*ie* ω B97XD, M06-2X and CAM-B3LYP) were considered for key TS1, TS2 and INT2 structures with an $\Delta\Delta G^\ddagger < 2.5 \text{ kcal mol}^{-1}$. Product percentages were re-calculated to reproduce Table 1 (Appendix Table S5.12 and S5.15). Using a different set of energy did not change result patterns and the same conclusions can be drawn. The mean absolute error (MAE) of the calculated percentages compared to the experimental percentages were considered for various reaction systems at different of theory (Appendix Table S5.11, S5.13 and S5.16). In all cases, the calculated percentages with VRAI-selectivity have a noticeably lower MAE than those based on TST only.

Quasi-classical MD calculations are regarded as the gold-standard approach to confirm that dynamics effects are important in reaction selectivity. They are much more computationally demanding than the VRAI-selectivity approach. Therefore, MD simulations were carried out with Jprogdyn¹⁷⁵ with the lowest energy TS1 in the ketone-selective reaction with Cat-B. This is the simplest reaction system investigated in this study. Ten trajectories for up to 6 ps from the lowest energy TS, which took a month with the available computer systems. Each trial commenced from the chosen TS1, with trajectories going in direction that leads to the formation INT2 adduct. The MD calculations were terminated either upon the generation of a product or in the event of recrossing, characterised by the distance between forming C-C bonds exceeding 3.4 Å. Details of the calculation are in the Table 5.3.

Table 5.3. Results of the MD simulations

Run	Outcome	Timeframe of the simulation (fs)
1	Epoxide	555
2	Recrossing	1840
3	Recrossing	825
4	Epoxide	4855
5	Epoxide	1325
6	Recrossing	865
7	Epoxide	2250
8	Epoxide	2210
9	Epoxide	465
10	Epoxide	5330

The CPU time required to run the corresponding VRAI-selectivity calculation and the MD simulation for producing reaction trajectories were compared. With the DFT-optimised structures as a starting point, the VRAI-selectivity calculation took 3.1 seconds on MacBook Pro (2020, 2 GHz Quad-Core Intel Core i5). The scaled CPU time is 3.1. For MD simulations, running a reaction trajectory spanning 400 fs required 12 hours using 64 CUP units with icelake nodes (Intel Xeon scalable processors) on Cambridge CSD3 high-performance computing environment. The average timeframe for running one trajectory based on Table 5.3 is 2052 fs. The average scaled CPU time for running ten trajectories is 14 183 424. Therefore, MD calculations took four million times more CPU time from the identification of the TS than using VRAI-selectivity (*ie* $14\,183\,424 / 3.1 = 4\,575\,298$).

The outcome did not correspond to the experimental result. If there were a hundred times more computer power, more trajectories would be run, starting from a range of low-energy TS. This would probably give an outcome that corresponds to the experimental outcome but changes a demanding computational problem into an impractical one. VRAI-selectivity introduces some new approximations but gives answers consistent with experimental data almost instantly provided with the DFT calculation results.

In the analysis, steps are regarded as being controlled either by reaction dynamics or by TST. Even if the excess energy is large, however, partial dynamic control can occur, where some of the excess energy is dissipated to the environment and some by internal vibrational energy redistribution. The excess energy for the different processes is given in Table 5.4. A detailed analysis of the competition between these processes would require more data than currently available and is a topic for further study.

Table 5.4. ΔG values related to the dynamically controlled steps: ΔG from TS1 to INT2 ($\Delta G(\text{TS1} \rightarrow \text{INT2})$) and ΔG from INT2 to TS2 ($\Delta G^\ddagger(\text{TS2})$) of the major pathways are recorded below for each reaction. The major pathway is the pathway with the highest %(Pathway) in the corresponding calculation breakdown table (with VRAI-selectivity) of the reaction. The level of theory is $\omega\text{B97XD/6-311g(d,p)}/\text{B3LYP-D3/6-31g(d)}$. (Unit: kcal mol^{-1})

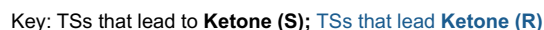
	$\Delta G(\text{TS1} \rightarrow \text{INT2})$	$\Delta G^\ddagger(\text{TS2})$
Ketone-selective reaction - Cat-B	-8.4	3.4
Ketone-selective reaction - Cat-C	-9.3	3.2
Ketone-selective reaction - Cat-D	-9.3	3.2
Epoxide-selective reaction - Cat-E	-11.6	1.2
Aldehyde-selective reaction - Cat-F	-7.4	1.6

Overall, the findings demonstrate the importance of considering both reaction kinetics and dynamics in predicting selectivity. A detailed breakdown of the calculations is given in Appendix to Chapter 5, 5.C.

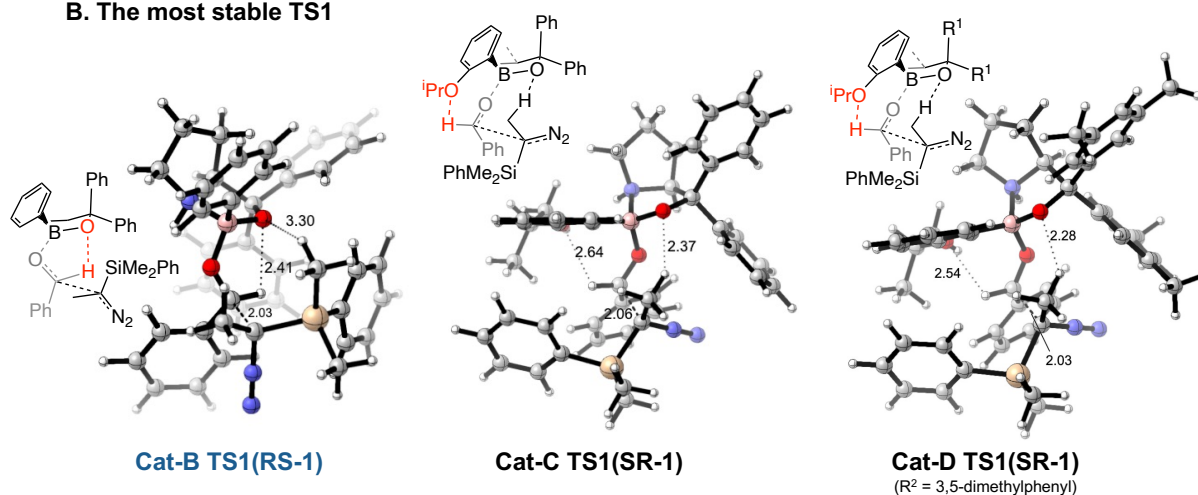
The results lead to some general implications on explaining the reaction selectivity.

This section focuses on the ketone-selective reactions with Cat-B, Cat-C and Cat-D. The enantioselectivity for these reactions exhibits significant variation with slight modifications in the COBI catalyst, specifically the addition of the oxy-isopropyl (iPrO) and methyl groups on the phenyl rings. As previously concluded, the stereochemistry for these systems is governed by TS1.

A. Energy distributes of key TS1: ketone synthesis reactions



B. The most stable TS1



C. Conformational differences between low-energy SR TS1 with Cat-C

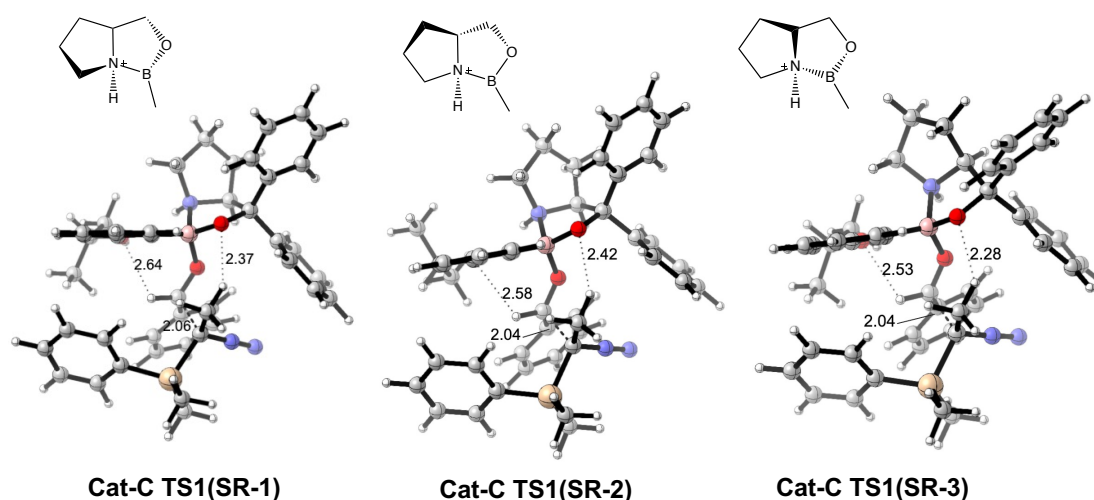


Figure 5.10. Key TS1s of ketone-selective reactions: A) An energy level diagram for key TS1s (with $\Delta\Delta G^\ddagger < 1.4$ kcal mol⁻¹) is presented at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) level of theory. For each TS1, the $\Delta G^\ddagger$ in kcal mol⁻¹ and stereochemistry at C4 and C5 centre are labelled. $\Delta G^\ddagger$ values are calculated relative to the lowest energy corresponding INT1 and diazo complex. The TS1 with SR and SS stereochemistry in reaction with Cat-B and TS1 with RR and RS stereochemistry in reaction with Cat-C and Cat-D have a $\Delta\Delta G^\ddagger > 1.4$ kcal mol⁻¹. B) The structure of the TS1 with the lowest $\Delta G^\ddagger$: The bond length values of key H-bonding interactions and the forming C-C bond are labelled in the diagram. The structures on the diagram are named according to the C4 and C5 stereochemistry and numbered based on their energy ranking. The orientations of both the aldehyde and the diazo are inverted by the possibility of a hydrogen bond to the ortho ⁱPrO group. C) Structures of low-energy SR TS1 with Cat-C: The TSs are very similar conformationally. The subtle differences are at the fused ring and rotation of the ⁱPrO group.

The energy differences between the key TS1s are due to multiple factors besides the key H bond interactions. The dispersion interactions and distortions of the structure during the C-C bond formation were considered. Distortion-interaction analyses were carried out on key TS1 structures following the below procedure:

1. The xyz coordinates of the INT1 and diazo compound are extracted from the optimised TS1 structure. Single point energy calculations were conducted on the distorted INT1 and diazo structure at the ω B97XD/6-311g(d,p) level of theory. The U of the distorted structures were compared to the U of the ground state structures to obtain the distortion energy for INT1 and diazo individually, ie the $\Delta U_{d,INT1}^\ddagger$ and $\Delta U_{d,diazo}^\ddagger$.

2. Deduce the total distortion energy, $\Delta U_d^\ddagger$ (eq. 5):

$$\Delta U_d^\ddagger = \Delta U_{d,INT1}^\ddagger + \Delta U_{d,diazo}^\ddagger \quad (5)$$

3. Deduce the activation energy, $\Delta U^\ddagger$ (eq. 6):

$$\Delta U^\ddagger = U_{TS1} - (U_{\text{ground state,INT1}} + U_{\text{ground state,diazo}}). \quad (6)$$

U_{TS1} , $U_{\text{ground state,INT1}}$ and $U_{\text{ground state,diazo}}$ are U of the TS1, ground state INT1 and diazo structure respectively.

4. Deduce the total interaction energy, $\Delta U_i^\ddagger$ (eq. 7):

$$\Delta U_i^\ddagger = \Delta U^\ddagger - \Delta U_d^\ddagger \quad (7)$$

The calculated $\Delta U^\ddagger$, $\Delta U_d^\ddagger$ and $\Delta U_i^\ddagger$ are reported in Table 5.5 (Cat-B systems) and Appendix Table S5.17 (Cat-C and Cat-D systems). The level of theory for all U values is ω B97XD/6-311g(d,p).

Stable TS1 structures tend to exhibit a more pronounced degree of structural distortion compared to the ground state as well as stronger non-covalent interactions between the diazo and INT1 complex. The results echo with the strong positive correlation between the $\Delta\Delta G^\ddagger$ of TS1 and the length of the forming C-C bond across reactions with Cat-B, Cat-C and Cat-D (Figure 5.11). Low-barrier C-C bond-forming pathways tend to have an earlier TS1.

Table 5.5. Distortion-interaction analysis results for the ketone-selective reaction (Cat-B). The level of theory for all values reported below is ω B97XD/6-311g(d,p). (Unit: kcal mol⁻¹)

Stereochemistry	$\Delta U^\ddagger$	$\Delta U_d^\ddagger$	$\Delta U_i^\ddagger$	$\Delta\Delta U_d^\ddagger$	$\Delta\Delta U_i^\ddagger$
RR-1	33.3	89.2	-55.9	6.7	-10.9
RS-1	33.3	88.5	-55.2	6.0	-10.2
SR-1	35.9	83.0	-47.1	0.5	-2.1
SS-1	37.5	82.5	-45.0	0.0	0.0

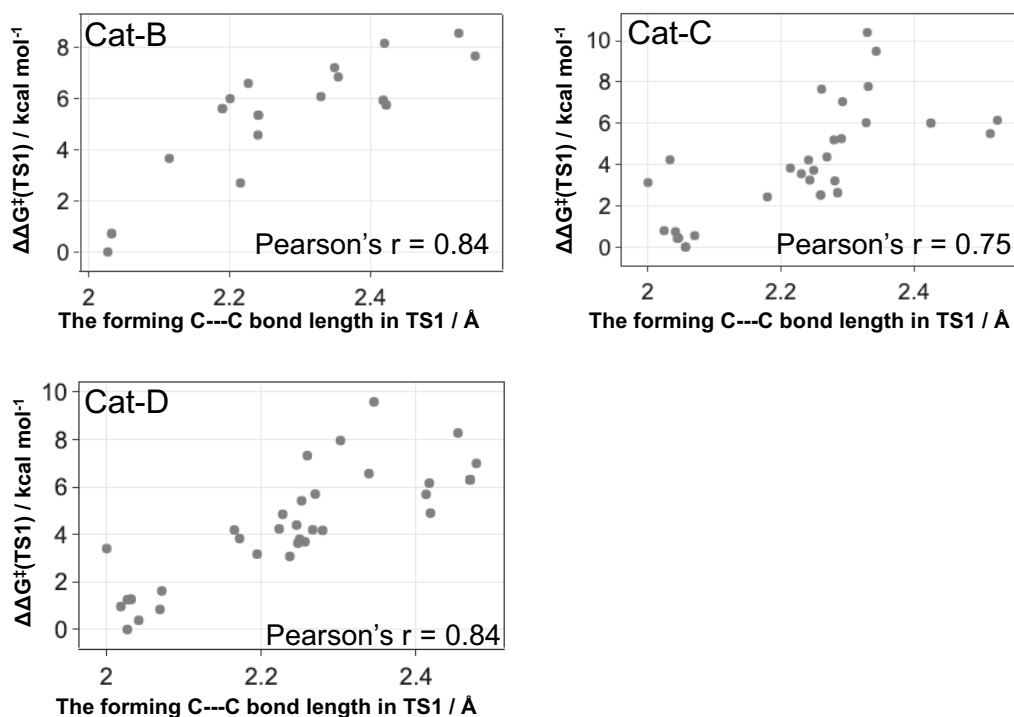


Figure 5.11. The ketone-selective reaction with Cat-B, Cat-C and Cat-D: $\Delta\Delta G^\ddagger$ of TS1 vs the forming C—C bond length. Level of theory: ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d)

Reactions with Cat-C and Cat-D are differed by the methyl groups at position 3 and 5 of the R¹ phenyl rings, which change the energy distributions of key TS1 structures (Figure 5.10.A). Firstly, $\Delta G^\ddagger$ of the TS1s in the reaction with Cat-D is slightly lower compared to the reaction with Cat-C. Secondly, $\Delta\Delta G^\ddagger$ difference between the lowest energy *SR* TS1 and *SS* TS1 increases from using Cat-C to Cat-D, which effectively reduces the percentage of the minor *R* ketone product. Thirdly, VRAI-selectivity analyses on the process beyond TS1 show changes in the PES upon the addition of methyl groups. Across pathways via the low-energy *SR* TS1s, the *S* ketone is more dynamically favourable compared to *R* aldehyde and *SS* epoxide in the reaction with Cat-D compared to Cat-C.

5.5.2 Chemoselectivity

This section compares the epoxide and aldehyde-selective reaction with a focus on chemoselectivity. Ryu *et al.* have commented that the electron-withdrawing group substituted aldehyde is in favour of the pathway to epoxide.¹⁵⁴ Besides the aldehyde substrate, the chosen epoxide and aldehyde-selective reaction only differ by alkyl substituents on the phenyl groups in the catalyst and diazo substrate.

Introducing the electron-withdrawing carbonyl group on the aldehyde substrate changes the PES significantly (Figure 5.8.A and 5.9.A). The pathway to epoxide is kinetically unfavourable in the aldehyde-selective reaction and higher in free energy than the corresponding TS1. TS2 structures to the epoxide have a $\Delta G^\ddagger$ of at least 6.72 kcal mol⁻¹, which corresponds to a $\Delta\Delta G^\ddagger = 5.16$ kcal mol⁻¹ when compared to the lowest energy TS2 structure. In the epoxide-selective reaction, the $\Delta G^\ddagger$ for TS2 structures to the epoxide are approximately 1 kcal mol⁻¹. On the other hand, the corresponding TS2s to aldehyde and ketone have comparable kinetic barriers with a $\Delta G^\ddagger$ in the range of 1.2-3.5 kcal mol⁻¹ in both the epoxide and aldehyde-selective reaction. The change in the energy barrier of the epoxide pathways results in the process being dynamically controlled and lead to different chemoselectivities. Hirshfeld charge analyses were conducted and studied the charge distribution by functional groups on key TS structures (Figure 5.12 and 5.13). The differences in charge distribution are insignificant across TS2s from both reactions. However, in TS1(SS-1) of the epoxide-selective reaction ($\Delta G^\ddagger = 4.43$ kcal mol⁻¹), the COPh group on the aldehyde withdraws negative charges and is noticeably less positive compared to the corresponding Ph group in the TS1(RS-1) of the aldehyde-selective reaction ($\Delta G^\ddagger = 7.59$ kcal mol⁻¹). The chemoselectivity of the epoxide-selective reaction is dynamically controlled by the trajectory from TS1 and the presence of the COPh group affects the charge distribution in TS1.

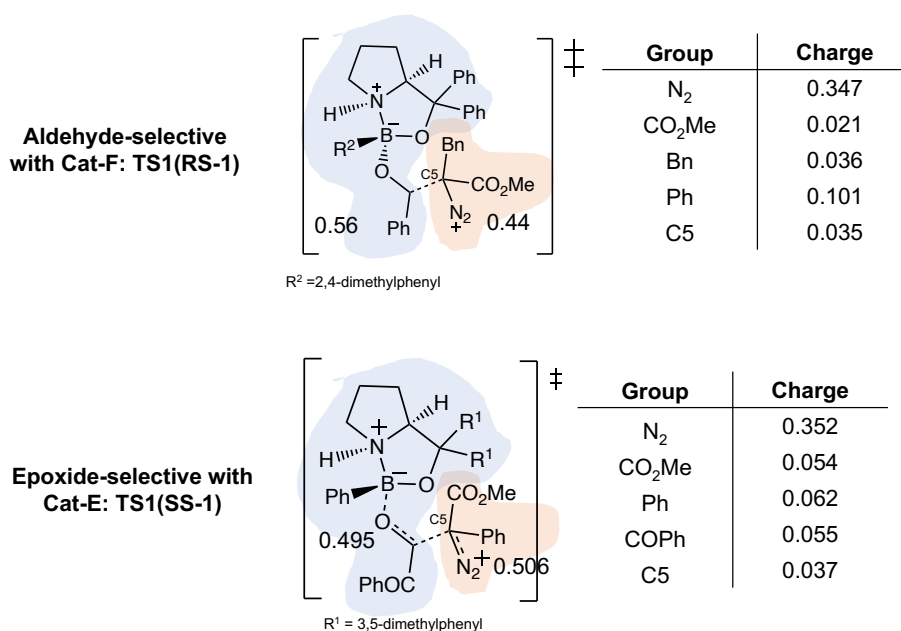
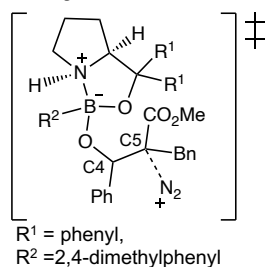


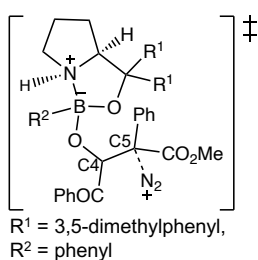
Figure 5.12. Hirshfeld charge analyses on TS1 structures for the aldehyde and epoxide-selective reaction. Pathways linked to the above TS1 contribute more than 50% of the final product composition. (Level of theory: ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d))

Aldehyde-selective with Cat-F:



Stereochemistry	RS Aldehyde	RS Ketone	RS Epoxide
Product			
N ₂	0.327	0.326	0.341
CO ₂ Me	0.077	0.079	0.080
Bn	0.089	0.146	0.123
Ph	0.066	0.028	0.016
C5	0.157	0.149	0.145

Epoxide-selective with Cat-E:



Stereochemistry	SS Aldehyde	SS Ketone	SS Epoxide
Product			
N ₂	0.398	0.311	0.341
CO ₂ Me	0.072	0.068	0.093
Ph	0.149	0.197	0.170
COPh	0.063	0.113	0.036
C5	0.119	0.123	0.123

Figure 5.13. Hirshfeld charge analyses on TS2 structures for the aldehyde and epoxide-selective reaction. The chosen TS2 structures have the lowest $\Delta G^\ddagger$ among other TS2 structures that lead to the same product. ‘Stereochemistry’ refers to the stereochemistry at position C4 and C5.

(Level of theory: ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d))

5.6 Summary

The investigation of the COBI-catalysed reactions suggests that reaction dynamics play a significant role in controlling selectivity. In the ketone and epoxide-selective reactions, the stereochemistry is controlled kinetically by the initial addition between the diazo and the catalyst-aldehyde complex (TS1). However, the process beyond TS1 is controlled by reaction dynamics, which influence the chemoselectivity of the outcome. In the aldehyde-selective reaction, both the stereoselectivity and the chemoselectivity are dynamically controlled.

Some general implications on the reaction selectivity were also concluded. Modifying the substituents on the substrate or catalyst leads to changes in the PES. In the ketone-selective reaction systems, the addition of an ⁱPrO group on the catalyst reverses the orientation of the aldehyde by providing the opportunity to form a new, favourable hydrogen bond. This leads to a reverse in enantioselectivity. Changing the phenyl rings to 3,5-dimethylphenyl favours the major product (*ie* *S* ketone) both chemo- and stereo-selectively. CPh-substituted aldehyde lowers the reaction barrier beyond TS1, which favours the synthesis of epoxides.

These results highlight the need for a more thorough exploration of conformational space and demonstrate the effectiveness of the VRAI-selectivity algorithm in predicting outcomes of dynamically controlled processes. This study provides an important step towards a deeper understanding of the complex interplay between reaction dynamics and selectivity in organic synthesis. Moving forward, the work in this chapter encourages the computational organic chemistry community to consider the procedure presented in this paper, which will provide a more comprehensive and accurate understanding of chemical reactivity to drive the development of better catalyst and synthesis strategies.

Chapter 6

Predicting sites of reaction based on chemistry within two bonds

6.1 Background

This chapter focus on understanding the possibilities and limits of data-driven methods for predicting molecular reactivity.

The growth in the application of machine learning models in predicting chemical reactivity has brought numerous challenges in model construction to light.^{62,95,97,176,177} One such challenge is formulating representations of chemical reactions or reacting molecules for machine learning.^{178–180} The optimal representations should be interpretable by computers while covering the traits of the chemical system relevant to the target property to be predicted by the model. The most common approach involves treating each reaction or set of reacting molecules as a single entity, where the descriptors are derived or calculated at the molecular level.^{181,182} For example, multiple fingerprint features are commonly used in machine learning for their robustness and applicability in dealing with a wide range of chemistry problems.¹⁸³ Most recent developments in the field of research are to adapt large language models for predicting reactivity with text-based representations as input.^{91,92} Computed and experimental physical and chemical parameters that quantify electric and steric factors have also proven effective.⁹³ Noteworthy successes include predicting the yield of C-N cross-coupling by Doyle *et al.*¹⁸⁴ and determining the enantioselectivity of phosphoric acid-catalysed reactions by Denmark *et al.*¹⁸⁵ Often, parameters of key atoms involved in the reactions are also included as part of the descriptor array. Using molecular-level representations necessitates the availability of extensive datasets, ideally comprising thousands of reactions. This not only introduces the challenge of data scarcity but also underscores the need for upholding quality during data gathering.^{94,186,187} The development of Chematica from Grzybowski *et al.* and Merck also show that AI-driven methods alone do not seem to work as well as hand-derived rules.^{188,189} Why not? What are the limits of pure machine learning methods? This will be investigated in this work.

Here, a different approach has been taken to predicting chemical reactivity. Atomistic approaches, or atomic fingerprint representations, have been explored in the construction of machine-learning potentials and for treating inorganic lattice structures.^{190–193} Within organic chemistry, Jensen *et al.* has utilised atomic-based descriptors for predicting the regioselectivity of electrophilic aromatic substitution reactions.¹⁹⁴ This project is founded on the hypothesis that the reactivity of an atom can be predicted if there are adequate knowledge about the atom and its local environment within the molecule, specifically within a range two bonds to the atom. This is a simplistic approach. The approach is investigated not because it will explain all of chemistry, but because it allows us to find out how far it can get. Here, machine learning models are trained on descriptors and labels designed for the individual atoms within

a molecule. Such an approach allows the exploration of data-driven methods on a relatively small dataset. Even with a dataset containing only a hundred reactions, the number of descriptors is likely to be in the thousands as it is the product of the number of atoms and the number of reactions. Evaluations were conducted on various datasets with varying complexity and diversity in chemical space. This allowed us to discover the extent to which this simple model had the capacity to encompass a substantial portion of reactivity within the chemical space of three types of reaction datasets. The outcome of the investigation, models for predicting sites of reaction, is useful for providing insights into the molecule reactivity and hinting at the possible transformations.

6.2 Data

6.2.1 Data gathering

Datasets of elementary organic reactions were collected and represented in the format of reaction SMILES with only the reactant and product component (*ie* reactant SMILES >> product SMILES). The chemical space of the dataset was restricted to reactions involving non-metal main group elements only. The gathered reaction datasets were categorised into three different types (Figure 6.1, 6.2 and Table 6.1): A) textbook reactions, B) reactions of a specific class and C) reaction datasets with diversity.

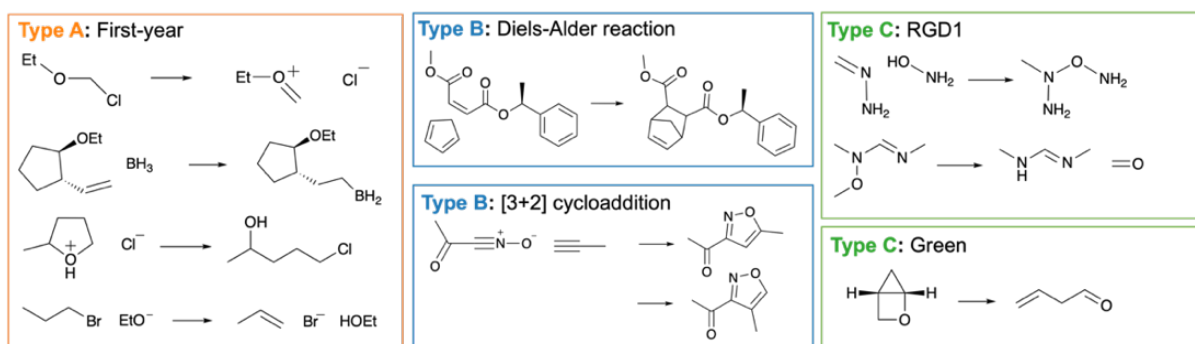


Figure 6.1. Examples of reactions from each dataset.

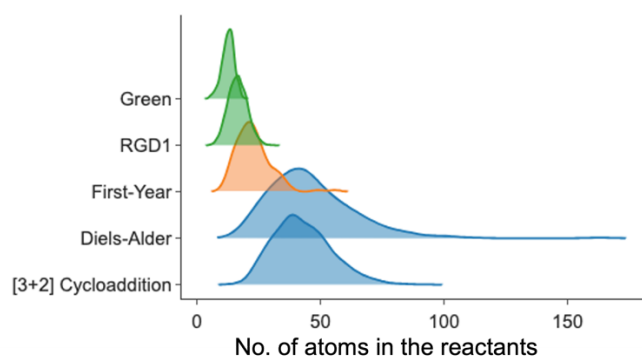


Figure 6.2. The distribution of the size of the chemical systems in each dataset.

Table 6.1. Overviews of the datasets after the data processing in this study.

Type	Dataset	No. of reactions	No. of sets reactions with the same reactants (<i>ie</i> competitive pathways)	Average no. of atoms in the reactants	Average no. of non-H atoms in the reactants
A	First-year	147	108	23	10
B	[3+2] cycloaddition	5953	2869	43	23
B	Diels-Alder reaction	2140	2140	46	23
C	RGD1	11281	10428	17	8
C	Green	321	252	13	7

Type A reaction datasets were manually generated based on the contents of the first-year organic chemistry lecture course at the University of Cambridge.¹⁹⁵ The first-year dataset was directly handpicked from the lecture handouts. The dataset comprises 146 reactions in total, including nucleophilic substitution, nucleophilic addition, elimination, enolisation and proton transfer (Figure 6.3).

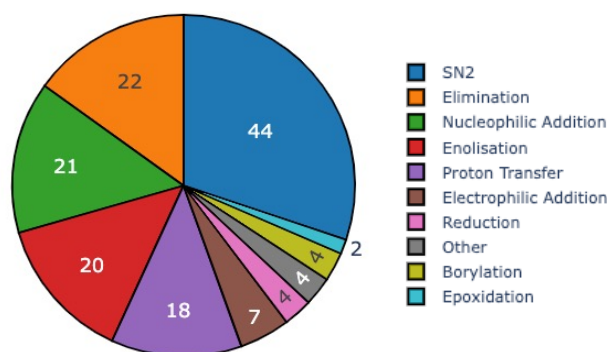


Figure 6.3. Composition of first-year dataset in pie charts

Type B reaction datasets comprise the computational dataset of [3+2] cycloaddition generated by Coley *et al.*¹⁹⁶ and the Diels-Alder reaction dataset collected by Tang *et al.*¹⁹⁷ All the [3+2] cycloaddition reactions and the first 2200 reactions from the original Diels-Alder dataset were taken for data processing.

Type C reaction datasets include the Reaction Graph Depth 1 (RGD1)¹⁹⁸ dataset from Savoie *et al.* and the elementary chemical reactions dataset from Green *et al.* (referred to as ‘the Green dataset’ below).¹⁹⁹ These datasets were generated in an automated fashion with programmed methods. RGD1 dataset uses the graphically defined elementary reaction step method to explore the chemical space and enumerate the reactions.^{200,201} The Green dataset relies on the growing string method to explore the potential energy surface and find the reaction pathways.^{53,77,79} For this investigation, both datasets were filtered to ensure that the reactions are thermodynamically favourable (*ie* $\Delta G < 0$ kcal mol⁻¹) with a low kinetic barrier (*ie* $\Delta G^\ddagger < 40$ kcal mol⁻¹).

6.2.2 Data processing

All the reaction SMILES strings collected in the data gathering stage were processed and standardised to the same format.

Atom-to-atom mapping was carried out on reactions in the first-year and Diels-Alder reaction datasets. The atoms in reactions of other datasets (*ie* the [3+2]-cycloaddition, RGD1 and Green dataset) have been mapped, where the reaction SMILES strings already contain the atom indexing. RXNMapper,²⁰² a transformer neural network model explicitly designed for atom-to-atom mapping, was employed to map non-hydrogen atoms. Subsequently, the hydrogen atoms were mapped based on the mapping of

the non-hydrogen atoms. As all reactions are elementary, it was assumed that no more than one hydrogen atom has changed its connectivity. The atom mapping from RXNMapper is not always perfect but has achieved the best accuracy among other methods in a recent benchmarking study.²⁰³ Schwaller *et al.* reported an 85% accuracy for RXNMapper,²⁰² tested on the USPTO data. Therefore, all the mapped first-year reaction SMILES strings were checked. Three out of 148 reactions had mapping errors, which were subsequently corrected manually. For the Diels-Alder reaction datasets, the mapping results in 100 reactions were manually checked. Errors were found in eight out of 100 reactions. Thus, it was assumed that errors are presented in 8% of the reactions in the Diels-Alder dataset.

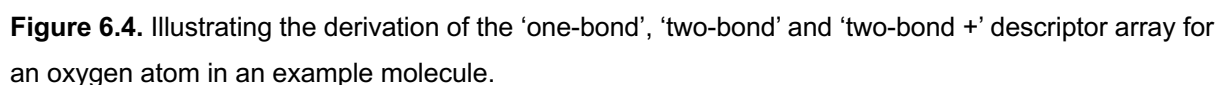
The sequential steps described below were executed on each dataset individually. Reactions with the same reactants were grouped together in recognition of competitive pathways. Within each set of reactions sharing the same reactants, the atoms were renumbered to ensure consistent atom numbering while considering the atom-to-atom mapping result. Additionally, a few reactions contain molecules that cannot be processed by the default setting of AllChem.EmbedMolecule in RDKit to generate 3D coordinates were also filtered out from each dataset. See Table 6.2 for the numerical details.

Table 6.2. Numerical records on filtering reactions that contain molecules that cannot be processed by the default setting of AllChem.EmbedMolecule in RDKit to generate 3D coordinates.

Dataset	Number of reactions before filtering	Number of reactions after filtering	Number of filtered reactions
First-year	147	147	0
[3+2] cycloaddition	5973	5953	20
Diels-Alder	2200	2140	60
RGD1	11406	11281	125
Green	332	321	12

In this project, the model is built on the assumption that a comprehensive understanding of an atom and its local chemical environment enables the prediction of its reactivity. Thus, the descriptors and labels are based on individual atoms.

Various descriptor array compositions have been considered, specially the 'one-bond', 'two-bond' and 'two-bond +' descriptor array (Figure 6.4).



- **Atom nature:** This includes the group and period of the corresponding element and specifies whether the atom belongs to one of the key organic elements (*ie* H, C, N, O, B, F, Cl, Br, Si, P or S).
- **Self-connectivity:** This part of the descriptor array specifies the number of key organic element atoms among its neighbours and the total number of neighbouring atoms.

- All features in the 'one-bond' descriptor array.

- Neighbours' connectivity: In accordance with the assumption that an atom has a maximum of 4 neighbours, this part of the descriptor array incorporates the self-connectivity descriptor arrays of the neighbouring atoms.

A 'two-bond +' descriptor array includes:

- All features in the 'two-bond' descriptor array.
- Bond strength: The bond strength component involves categorising the bonds to which the atom is connected into one of the 86 bond classes across 12 bond types (*ie* CC, CN, CH, CCl, CO, CS, CF, NO, HN, HO and OP). The complete list of bond classes is in Appendix Table S6.2. The classification of the 86 bond classes was established using the method employed in the MolE8 machine-learned potential energy surface model.²⁰⁴ The classification of bonds is based on the bond length of the specific bond type. The parametrisation procedure was repeated and the bond length distributions of each bond type on MMFF-optimised structures were studied using a dataset of 100 000 molecules from ChEMBL-28.²⁰⁵ Kernel density estimations (KDE) were applied to the histogram of the bond length distribution (Figure 6.6). The minima observed on the KDE curve were identified as the boundaries dividing each bond class (Appendix Table S6.2).

The bond strength descriptors explicitly provide information on the chemical environment of the atom beyond the two-bond range. The bond strength descriptors can distinguish between the same bond (*ie* same bond type and same bond order) but located in different chemical environments due to subtle bond length differences. For example, the single C-C bond within a conjugated system and in a saturated chain have slightly different bond lengths and are categorised into two bond strength classes (Figure 6.5).

- Rings: atoms that are in rings are marked, and the ring size is recorded if it is smaller than eight.

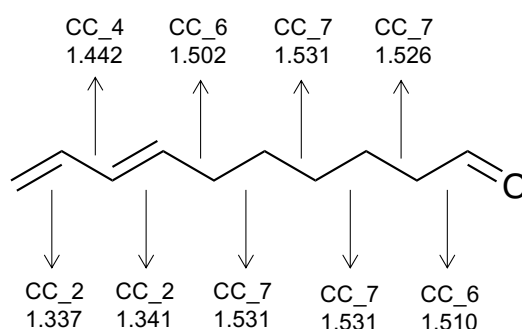


Figure 6.5. Bond strength class and bond length in Å for all the CC bonds in an example molecule.

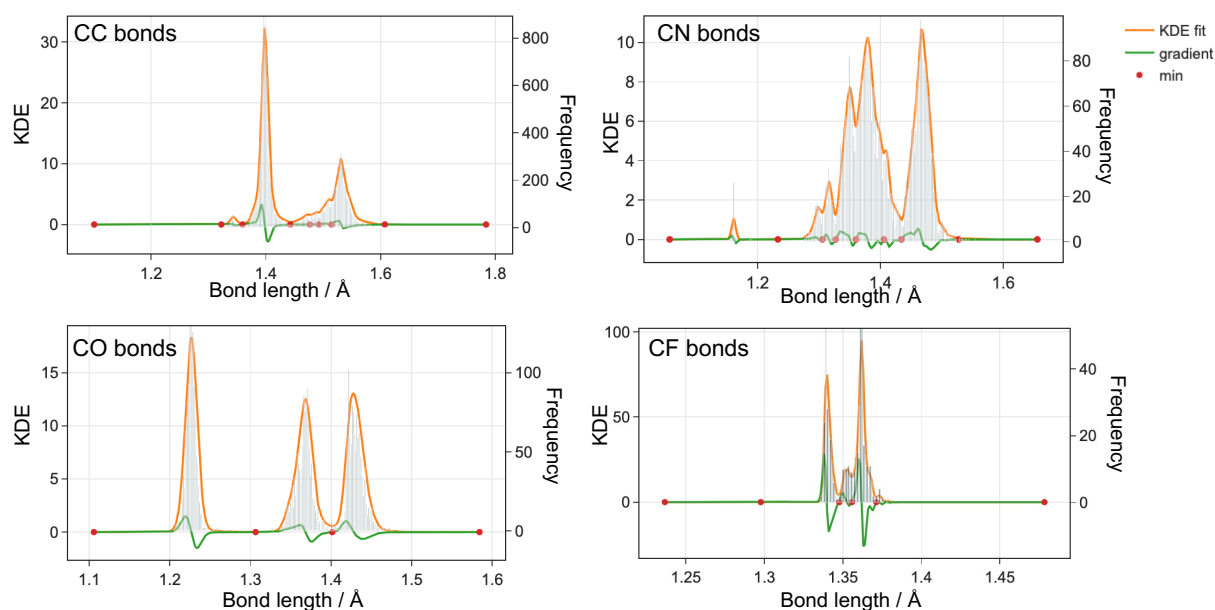


Figure 6.6. Derivation of the bond strength classes from bond length distribution histograms. The corresponding plot for CC, CN, CO and CF are given above as examples.^{204,205}

6.3.2 Label

The atoms are either labelled as 'reactive' or 'unreactive' (Figure 6.7).

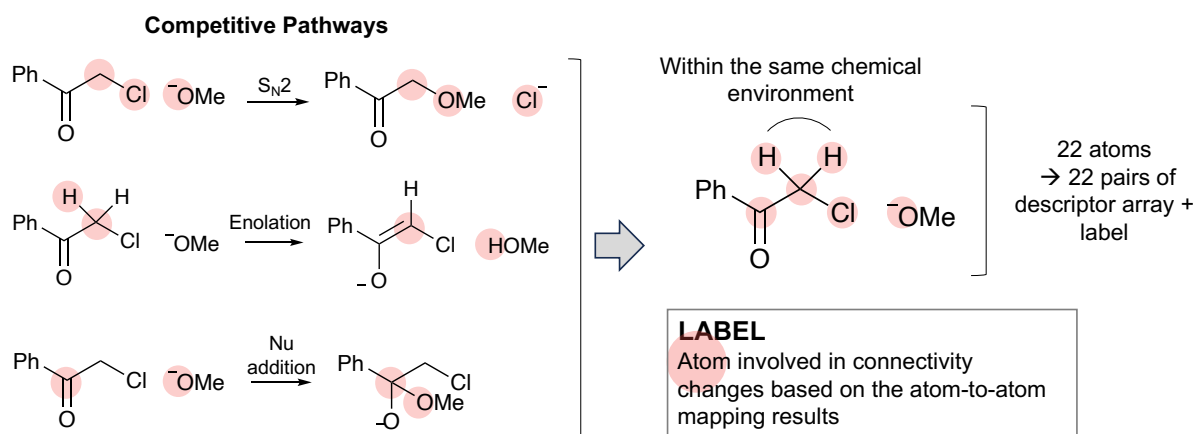


Figure 6.7. Illustrating the derivation of the label based on the atom-to-atom mapping result. An atom is considered reactive if it undergoes a connectivity change in one of the possible competitive reactions. Atoms within the same chemical environment were assigned identical labels.

To derive the label, the connectivity difference of each atom in the reactant and product was compared for every reaction through comparing the adjacency matrices based on atom-to-atom mapping results. Here, changes in connectivity refer to changes in the neighbouring atoms only. Changes in bond order have not been accounted. In accounting for competitive reactions from the same reactants, an atom is considered reactive if it undergoes a connectivity change in one of the possible competitive reactions.

The example in Figure 6.7 shows three competitive reactions that share the same reactants. There are 22 atoms in the reactants. Thus, this set of reactions would yield 22 pairs of a descriptor array and a corresponding label on reactivity.

Atoms within the same chemical environment were assigned identical labels. In Figure 6.7, both protons are likely to be extracted in the enolisation reaction, although only one of them will be involved in the actual reaction. Thus, for consistency, if one of the atoms within a chemical environment is found to be reactive, all other atoms within the same environment are regarded as reactive.

The chemical environment of atoms within a molecule was determined via *in-silico* isotopic labelling. For example, let us consider two atoms from the same chemical environment in a molecule. Two copies of RDKit⁴² Chem.Mol objects are generated from the molecule, followed by isotopic labelling on each atom individually in each Chem.Mol objects. Subsequently, by converting the Chem.Mol objects to InChI⁸⁹ representations, the two resulting InChI strings are expected to be identical.

6.3.3 Algorithm

For the choice of the algorithm, benchmarking studies have been carried out on various classification algorithms with the 'two-bond +' descriptor arrays, including random forest (RF), K-nearest neighbour, support vector, gaussian process and multi-layer perceptron classifier. The default setting from the scikit-learn package were used.¹³² While similar results in terms of accuracy were achieved across the different models, the models based on RF classifier showed slightly better performance than others (Appendix Table S6.3). In the text below, all the reported models are based on the RF classifier algorithm.

6.4 Results and Discussion

Model training and evaluation were performed for each dataset with different descriptor compositions (ie the 'one-bond', 'two-bond' and 'two-bond +' descriptor array composition). The details of the trained models and results from the evaluation are presented in Table 6.3.

Relatively small datasets were employed deliberately in model training to illustrate the effectiveness of the atomistic approach. The first-year dataset covers a diverse range of reactions (Table 6.3). The 30 sets of reactions were randomly selected from the dataset for testing, leaving 78 sets of reactions for training (Table 6.3 Entry 1-3). The compositions of the training and testing datasets were checked to ensure that all types of reactions present in the testing dataset have appeared in the training dataset. Compositions of the first-year training and testing dataset are presented in Appendix Figure S6.2. For all other datasets (ie [3+2] cycloaddition, Diels-Alder, RGD1 and Green; Table 6.3 Entry 4-9 and 13-18), 100 sets of reactions were randomly selected for training and testing, respectively. Combination of the datasets (Table 6.3 Entry 10-12 and 19-24) were also considered to assess the generalisability of the model.

The accuracy, precision and recall were reported on an atomistic basis. Precision is the number of true reactive predictions over the total number of reactive predictions. Recall is the number of true reactive predictions over the total number of reactive atoms in the dataset.

$$Precision = \frac{True\ reactive}{True\ reactive + False\ reactive} \quad (8)$$

$$recall = \frac{True\ reactive}{True\ reactive + False\ unreactive} \quad (9)$$

At the molecular level, the percentage sets of reactants with no fault predictions, no more than one fault prediction and all reactive atoms predicted correctly were calculated.

In Table 6.3, other than the models from the first-year dataset (Entry 1-3), only a fraction of the dataset was considered in training for models based on individual datasets (Entry 4-9 and 13-18). To validate the results, the following procedures were conducted on the [3+2] cycloaddition, Diels-Alder, RGD1 and Green dataset individually:

1. 100 sets of reactions with the same reactants were randomly selected for training and testing, respectively.
2. Model training was performed using the 'two-bond +' descriptor composition. The metrics from the evaluation were recorded after the testing.
3. The above steps were repeated ten times. The mean and standard deviation of the performance metrics were calculated.

The results from this random sampling test were recorded in Table 6.4. The mean values of the evaluation metrics are comparable to the results from Table 6.3. The standard deviations of the

performance metrics are also relatively small. This demonstrates the results from Table 6.3 (Entry 4-9 and 13-18) are representable.

Table 6.3. Performance of the models. Three different types of datasets (Type A, B and C) and the combinations of these datasets were considered in the model evaluation. The size of the dataset involved in training and testing is specified below. Let us take ‘First-year reactions: Train: 78 (105), Test: 30 (42) → total 147 reactions, ie 100% of the dataset’ as an example. ‘Train: 78 (105)’ implies that 78 sets of reactions with the same reactants, equivalent to 105 reactions, were involved in training. ‘Test: 30 (42)’ means 30 sets of reactions with the same reactants, equivalent to 42 reactions, were involved in testing. In total, 147 reactions were involved in the training and testing of the model. This corresponds to 100% of the first-year reaction dataset. For each set of data, ‘one-bond’, ‘two-bond’ and ‘two-bond +’ descriptor compositions were considered.

		% by atoms			% sets of reactants with:		
Entry	Descriptors	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
Type A							
First-year reactions: Train: 78 (105), Test: 30 (42) → total 147 reactions, ie 100% of the dataset							
1	One-bond	88.59%	73.64%	63.28%	10.0%	56.7%	43.3%
2	Two-bond	91.59%	83.96%	69.53%	26.7%	56.7%	56.7%
3	Two-bonds +	93.09%	87.27%	75.00%	40.0%	70.0%	63.3%
Type B							
[3+2] cycloaddition: Train: 100 (205), Test: 100 (219)→ total 424 reactions, ie 7% of the dataset							
4	One-bond	95.31%	92.16%	75.81%	25.0%	62.0%	35.0%
5	Two-bond	99.86%	99.77%	99.31%	96.0%	100.0%	97.0%
6	Two-bonds +	99.96%	99.77%	100.00%	99.0%	100.0%	100.0%
Diels-Alder: Train: 100 (100), Test: 100 (100) → total 200 reactions, ie 9% of the dataset							
7	One-bond	91.76%	90.43%	22.91%	0.0%	3.0%	0.0%
8	Two-bond	96.07%	88.74%	71.15%	27.0%	56.0%	37.0%
9	Two-bonds +	97.12%	90.20%	81.06%	44.0%	72.0%	60.0%
Combined: Train: 100 (205) [3+2] cycloaddition + 100 (100) Diels-Alder, Test: 100 (219) [3+2] cycloaddition + 100 (100) Diels-Alder							
10	One-bond	92.53%	85.11%	47.64%	13.0%	31.5%	14.5%
11	Two-bond	97.43%	93.56%	85.02%	56.5%	78.0%	64.5%
12	Two-bonds +	98.23%	94.08%	91.33%	68.0%	88.0%	80.5%
Type C							
RGD1: Train: 100 (107), Test: 100 (110) → total 217 reactions, ie 2% of the dataset							
13	One-bond	76.09%	65.29%	54.20%	6.0%	21.0%	24.0%
14	Two-bond	77.13%	65.00%	62.02%	5.0%	18.0%	25.0%
15	Two-bonds +	76.33%	63.10%	62.98%	6.0%	14.0%	28.0%
Green: Train: 100 (126), Test: 100 (120) → total 246 reactions, ie 76% of the dataset							
16	One-bond	82.66%	84.60%	76.38%	18.0%	37.0%	34.0%
17	Two-bond	84.07%	87.55%	76.38%	31.0%	46.0%	46.0%
18	Two-bonds +	84.24%	87.45%	76.95%	33.0%	49.0%	49.0%
Combined: Train: 100 (107) RGD1 + 100 (126) Green, Test: 100 (110) RGD1 + 100 (120) Green							
19	One-bond	78.24%	75.51%	62.92%	11.5%	29.0%	29.5%
20	Two-bond	79.29%	75.96%	66.25%	18.0%	33.0%	34.5%
21	Two-bonds +	79.79%	76.49%	67.30%	20.0%	35.0%	39.0%
Combining all datasets: the global model							
Train 78 (105) first-year + 100 (205) [3+2] cycloaddition + 100 (100) Diels-Alder + 100 (107) RGD1 + 100 (126) Green, Test: 30 (41) first-year + 100 (219) [3+2] cycloaddition + 100 (100) Diels-Alder + 100 (110) RGD1 + 100 (120) Green							
22	One-bond	87.00%	71.58%	45.22%	6.7%	19.5%	11.2%
23	Two-bond	91.37%	78.94%	70.55%	28.4%	44.7%	43.5%
24	Two-bonds +	92.50%	81.08%	75.69%	35.3%	50.7%	54.2%

Table 6.4. Performance for the random sampling test. Model training was performed using the ‘two-bond +’ descriptor composition.

Entry	Dataset	% by atoms			% sets of reactants with:		
		Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
Mean							
1	[3+2] cycloaddition	99.57%	98.75%	96.51%	82.8%	96.9%	86.2%
2	Diels-Alder	96.39%	88.24%	75.97%	40.60%	63.30%	57.40%
3	RGD1	78.01%	66.45%	55.01%	5.8%	19.5%	24.9%
4	Green	84.05%	86.80%	76.71%	27.1%	46.1%	42.6%
Standard Deviation							
5	[3+2] cycloaddition	0.36%	1.47%	3.24%	13.4%	4.2%	12.2%
6	Diels-Alder	0.55%	3.00%	3.68%	5.10%	5.19%	7.31%
7	RGD1	1.16%	2.55%	4.98%	1.9%	4.2%	5.7%
8	Green	0.78%	0.77%	1.99%	3.0%	3.6%	3.4%

6.4.1 Type A: the first-year dataset

The first-year dataset includes a variety of different reactions. Hence, it is a surprise that even a model trained with only the 'one-bond' descriptors can achieve an accuracy of 88.6% by atoms (Table 6.3 Entry 1). The good performance can be explained by looking at the dataset composition. Many first-year reactions, such as S_N2 reactions and nucleophilic addition reactions on carbonyl bonds, are quite simple and the sites of reaction only involve a few atoms. In these cases, the chemical environment within one bond to an atom is sufficient to determine the reactivity.

The model performance improves when the neighbours' connectivity and bond strength descriptors are included in the model training. With the 'two-bond +' descriptor array, there is no more than one fault prediction in 70% of the sets of reactants in the testing dataset (Table 6.3 Entry 3). This implies that the local chemical environment within two bonds of an atom is often sufficient to determine its reactivity. Figure 6.8 showcases examples where noticeable improvements are observed when extending the range of chemical environments included in the descriptors for each atom. Example A4 shows possible enolisation reaction pathways from an 1,3-dicarbonyl compound. Alpha carbon atoms are correctly predicted as reactive when including the neighbours' connectivity and bond strength descriptors in model training. Example A5 describes a scenario where primary carbon atoms, which tend to be involved in S_N2 or E2 reactions when connected to a good leaving group, are predicted incorrectly as reactive. These primary carbon atoms are predicted as unreactive in the model trained with the bond strength descriptors, which presumably reinforce the recognition that the carbon atoms are not connected to any good leaving group. The predictions on the reactivity can hint at the possible products from the reactants. In many cases, only the predictions from the 'two-bond +' model can account for all the competitive pathways in the dataset. These examples highlight the effectiveness of the bond strength and connectivity descriptors in describing the chemistry within a two-bond range and indicating the local chemical environment.

Model evaluation with the first-year dataset: Train: 78, Test: 30

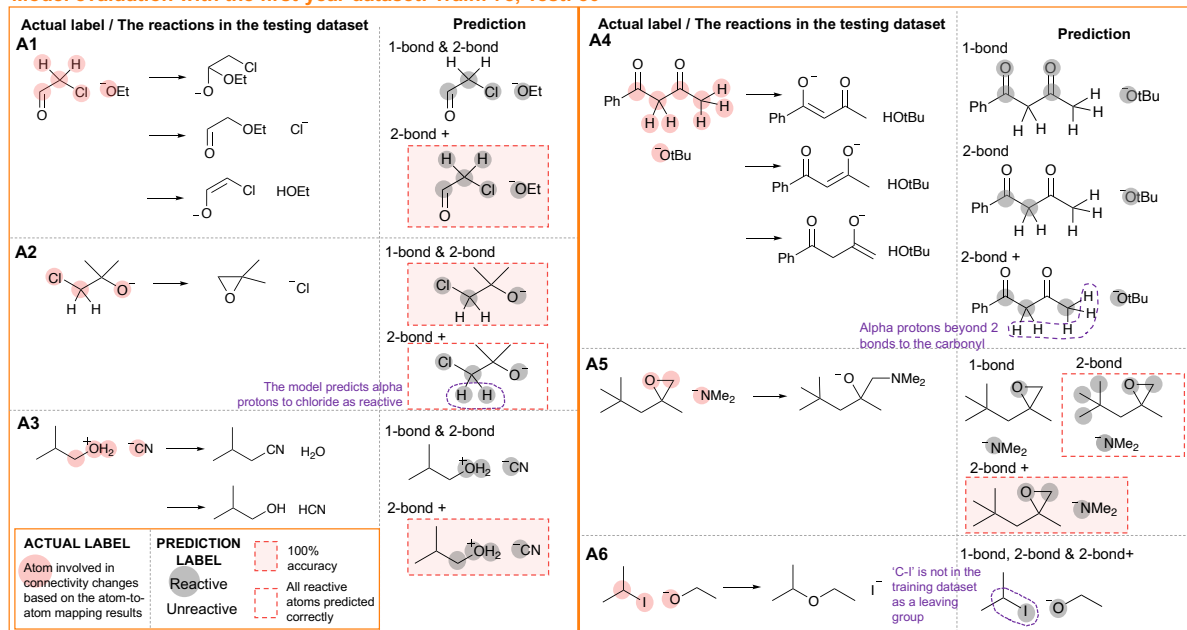


Figure 6.8. Case studies on the results from model evaluation with the first-year dataset. The annotations on the fault predictions are in purple.

Case studies also help to understand the limitations of the model (Figure 6.8). Firstly, it is within the expectations that the models cannot predict the reactivity of atoms which requires the consideration of the chemical environment beyond two bonds from the site of connectivity changes. While bond strength and the ring component descriptors recognise the different chemical environments, they do not explicitly describe the specific atom arrangements beyond the two-bond distance from the atom. Many of the false predictions of unreactive behaviour of the 'two-bond +' model come from the alpha proton of the carbonyl or leaving group in enolisation E2 reactions. The motif at the reaction site extends beyond two bonds for these reactions. An exception is example A1. The alpha protons in the aldehyde are accurately predicted as reactive even though the protons are three bonds away from the carbonyl group. Here, it is likely that the correct predictions are made based on the wrong reason. These protons are predicted to be reactive because they are at the alpha position relative to the chloride. Cases of mistaken reactive predictions, such as example A2, show that the model consistently predicted the alpha proton to the chloride group as reactive.

At times, the model may encounter challenges in extrapolating beyond the training data, even when there are reactions of the same type in the training dataset. This becomes especially apparent when dealing with unfamiliar chemical groups that involve changes in connectivity. Figure 6.8 A6 illustrates such a scenario. 'C-I' has not been presented as a leaving group in the training dataset S_N2 reactions, contributing to the fault unreactive prediction of the C atom in 'C-I' for example A6 in the testing dataset.

Tests on reactions from first-year exam questions

Actual label / The reaction from the exam			From reactant to the intermediate	Prediction (2-bond+)	Accuracy (to product)	Accuracy (to intermediate)			
A7								29/30	30/30
A8					59/60	60/60			
A9					28/29	NA			
ACTUAL LABEL Atom involved in connectivity changes based on the atom-to-atom mapping results			PREDICTION LABEL Reactive Unreactive						

Figure 6.9. Predictions from the ‘two-bond +’ model trained on all first-year reactions using reactions from first-year exams at the University of Cambridge. The accuracy by atom values was calculated based on the actual label derived from atom-to-atom mapping results compared to the product and the intermediate.

To assess the model trained on all first-year reactions (*ie* 2473 atomic descriptor arrays and labels from 108 sets of reactants), tests were conducted using past exam questions from the University of Cambridge. The outcome is presented in Figure 6.9. There is no more than one mistaken prediction in each reaction. This demonstrates the performance of the model in a real-world context.

6.4.2 Type B: the [3+2] cycloaddition and the Diels-Alder dataset

The model performs well in predicting the atom reactivity in reactants of reactions where there are many similar examples in the training dataset. In type B datasets, only a single type of reaction is presented, and therefore, good performances were achieved. The accuracy by atom for type A datasets, the [3+2] cycloaddition and the Diels-Alder dataset, are close to 100% for the 'two-bond +' model (Table 6.3 Entry 6 and 9). Example B1-B3 in Figure 6.10 are cases of successful predictions from the model evaluation. These cases also demonstrate the improvements in performance when introducing the bond strength and neighbours' connectivity descriptor components (*ie* moving from 'one-bond' to 'two-bond' and 'two-bond +'). Although the bond-forming carbon atoms are more than two bonds apart in the diene motif of substrates in Diels-Alder reactions, the bond strength descriptor components should be capable of distinguishing between the various carbon atoms within the diene motif. Figure 6.5 illustrates such a situation with a molecule containing a diene motif.

Taking away the bond strength descriptors has not significantly defected the performance of the models trained on type B datasets. This is out of the expectations. One possible explanation is that the connectivity descriptors also implicitly indicate about the chemical environment beyond the range of two bonds. For example, the total number of neighbouring atoms of atoms at a two-bond distance may indirectly hint at the hybridisation of the atoms three bonds away. The self-connectivity descriptor components also include the number of neighbouring atoms for its neighbours. Thus, even for the 'one-bond' model, there are no fault predictions in 25% of the sets of reactants in the evaluation with the [3+2] cycloaddition dataset.

The quality of the dataset also matters. In data processing, it was estimated that atom-to-atom mapping errors are presented in 8% of the reactions in the Diels-Alder reaction dataset. These errors lead to drop in performance of the models trained on Diels-Alder reactions. To verify this claim, tests were conducted on the 100 reactions for which mapping errors has been manually checked. Take-one-out cross-validations were performed on the 100 Diels-Alder reactions before and after the corrections of atom-to-atom mapping errors. Noticeable improvement has been observed in the accuracy, precision and recall value by atoms (Table 6.5).

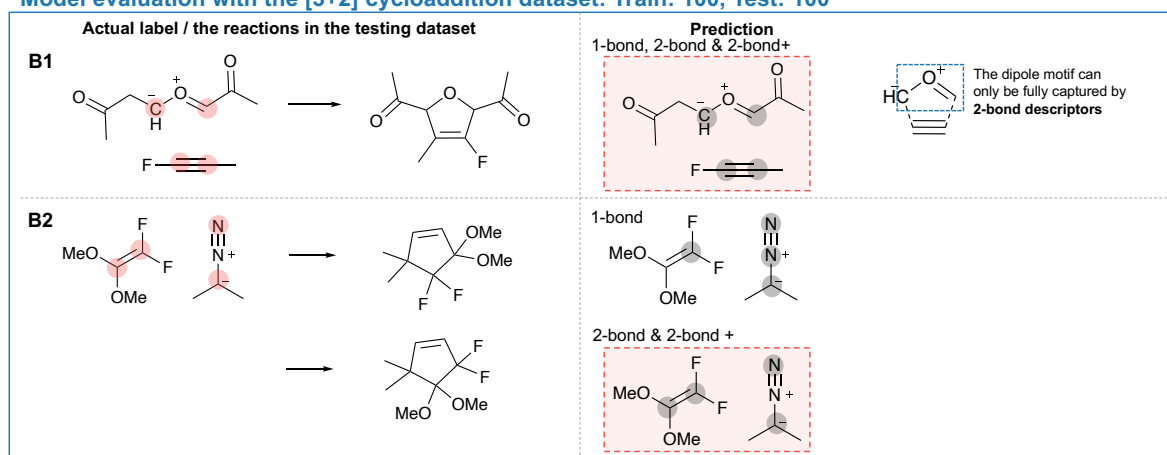
Another observation from the study on the Diels-Alder dataset is that the model may sometimes hint at potential competitive pathways leading to products not presented in the dataset (Figure 6.10 B4). For instance, in the B4 Diels-Alder reaction, the model also highlights an alternative potential dienophile position in the substrate, which is a reasonable pathway.

Table 6.5. Performance for the take-one-out cross-validation test. The unit for training and testing data is ‘set’, *ie* a set of competitive reactions with the same reactants. The ‘2-bond +’ descriptor composition was used. The training is performed with the RF classifier. The specific datasets and number of sets of reactions involved are specified below.

Entry	Dataset (no. of sets of reactions)	% by atoms			% sets of reactants with:		
		Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
1	Diels-Alder (100)	97.28%	90.56%	82.38%	47.0%	76.0%	63.0%
2	Diels-Alder (100)*	98.47%	92.21%	91.55%	58.0%	85.0%	72.0%

*After correcting the atom-to-atom mapping errors

Model evaluation with the [3+2] cycloaddition dataset: Train: 100, Test: 100



Model evaluation with the Diels-Alder reaction dataset: Train: 100, Test: 100

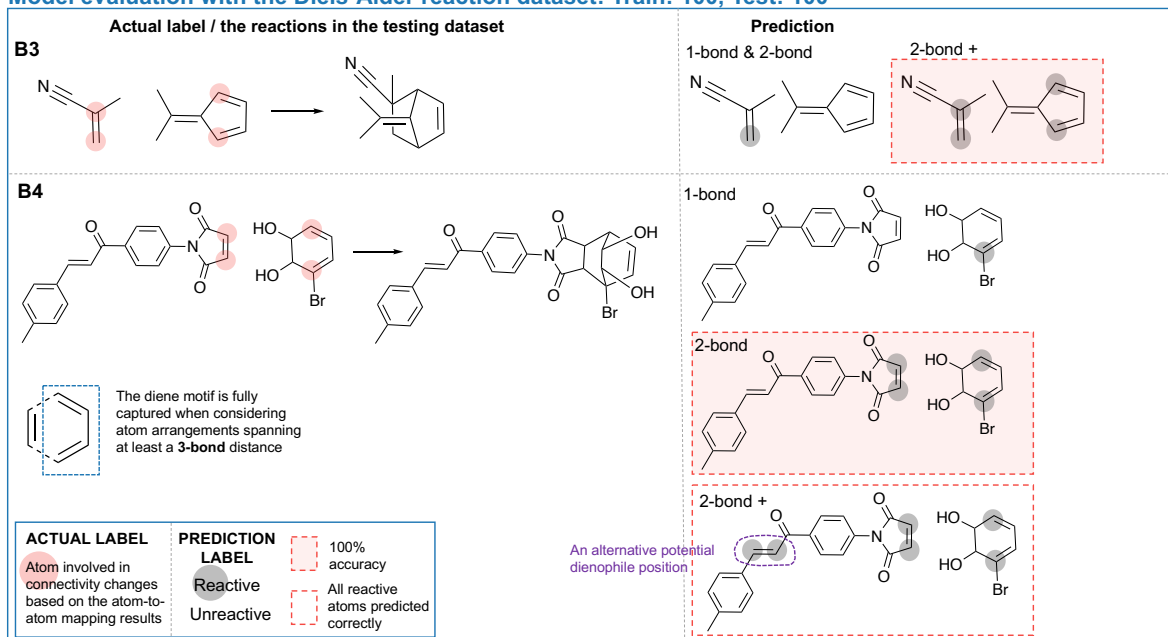
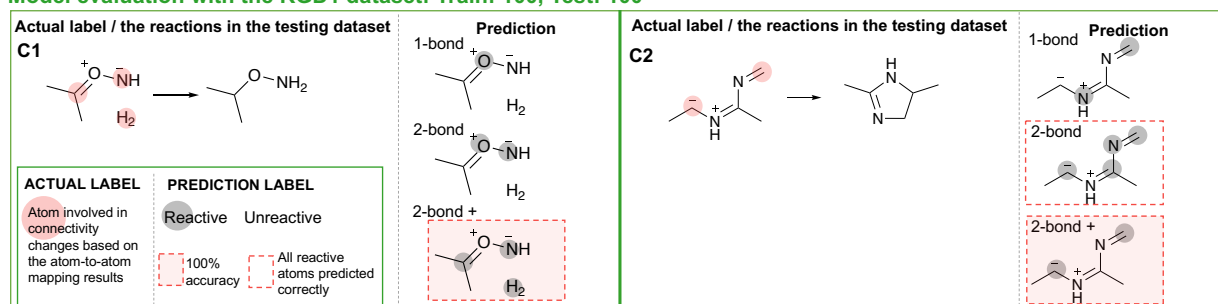


Figure 6.10. Case studies on the results from model evaluation with the type B datasets: the [3+2] cycloaddition¹⁹⁶ and the Diels-Alder reaction¹⁹⁷ dataset.

6.4.3 Type C: the RGD1 and the green dataset

The RDG1 and the Green datasets are more challenging and complex compared to type A and type B datasets. The reactions often involve multiple steps mechanistically and are not the typical textbook reactions. The RDG-1 and the Green dataset were analysed based on reaction templates to quantify the types of reactions present in the dataset. A reaction template in SMILES string was generated for each reaction, highlighting the motif at the reaction site. The motif at the reaction site covers the atoms involved in changes in connectivity. If less than two atoms involved connectivity changes, the non-H neighbouring atoms of the reactive atoms were also accounted in the template. 100 random reactions were selected from each dataset for the test. 82 and 59 unique templates were generated from 100 RDG1 and 100 Green reactions respectively. However, even for the challenging RDG-1 and Green datasets, the trained models on just 100 sets of reactions with 'two-bond +' descriptors gave a surprisingly decent performance of 76% and 84% accuracy respectively (Table 6.3, Entry 15 and 18), which is much better than random guessing (*ie* the accuracy of random guessing is 50%). The cases presented in Figure 6.11 are examples of the predictions from the model evaluation. It is important to emphasise the representativeness of the above results despite the small dataset selected for training and evaluation for the RGD1 dataset. Repeating the test on the RGD1 dataset with different sets of randomly selected reactions gives consistent results (Table 6.4). This implies that models trained from 1% of the dataset can already cover quite a substantial amount of chemistry in the RGD1 dataset.

Model evaluation with the RGD1 dataset: Train: 100, Test: 100



Model evaluation with the Green dataset: Train: 100, Test: 100

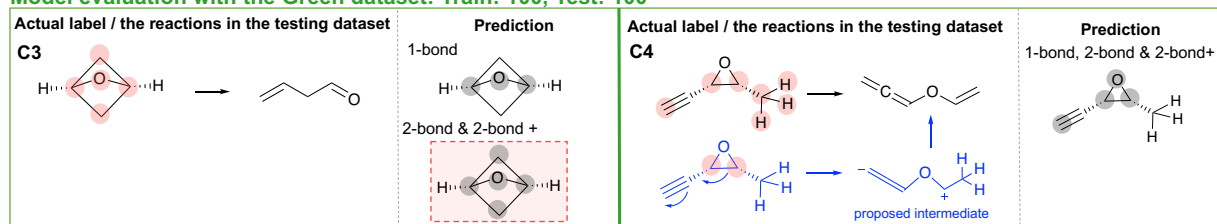


Figure 6.11. Case studies on the results from model evaluation with the type C datasets: the RGD1¹⁹⁸ and the Green¹⁹⁹ dataset. In C4, the prediction accounts for the connectivity changes from the reactant to the proposed intermediate.

It is also worth noting that the ‘one-bond’ models already exhibit decent performance. The enhancement in performance when incorporating ‘two-bond’ and ‘two-bond +’ descriptors is noticeable but less pronounced compared to the improvements seen in models trained with type A and B datasets. This observation can be explained by the fact that the chemical systems in RGD1 and Green reactions are relatively small. The average number of atoms in the reactants within type C datasets is below 17, in contrast to 23 in the first-year reaction dataset and exceeding 40 in the type B datasets (Table 6.1). The substructure covered by the ‘one-bond’ descriptors is often nearly half of the molecules in type C reactants. Thus, the models demonstrate satisfactory performance even without the additional ‘two-bond’ and ‘two-bond +’ descriptors.

Increasing the number of sets of reactions in the training dataset improves the performance of the ‘two-bond +’ RGD1 model significantly (Figure 6.12 and Appendix Table S6.4). 100 sets of reactions are equivalent to 1% of the RGD1 dataset. The model trained with 300 sets of RGD1 reactions (*ie* 3% of the dataset) has an accuracy of 81%. The model trained with nearly all of the RGD1 dataset, *ie* equivalent to 10,300 sets of reactions, has an accuracy of 84%.

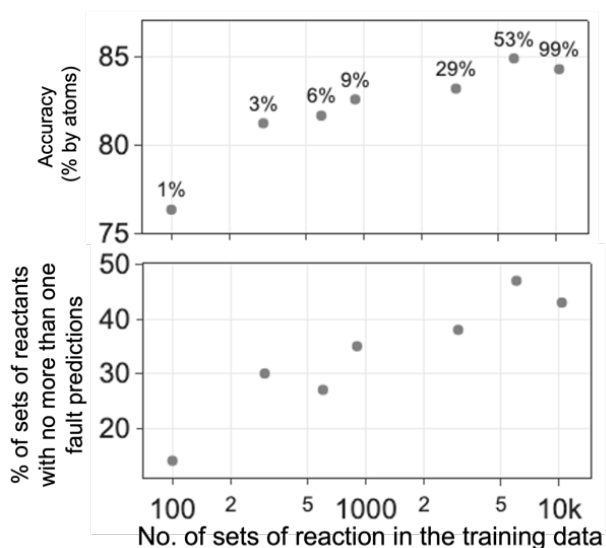


Figure 6.12. Increasing the size of the training data leads to improved performance metrics for the ‘two-bond +’ model trained using the RGD1¹⁹⁸ dataset. The percentage of data used in the training as part of the RGD1 dataset are labelled in the accuracy plot.

6.4.4 Model generalisability

Models trained from combinations of different datasets, specifically [3+2] cycloaddition + Diels-Alder, RGD1 + Green and all datasets together (Table 6.3 Entry 10-12 and 19-24), were also considered. In all three scenarios, an improvement in the performance has been observed when moving from the model trained on the 'one-bond' to the 'two-bond' and 'two-bond +' descriptor composition. The accuracy values from the combined dataset models are very similar to the average accuracy of models trained from the individual datasets. For example, the average accuracy for the 'two-bond +' models trained from the individual datasets (Table 6.3 Entry 3, 6, 9, 15 and 18) is 90.15%. The accuracy of the model trained from the combination of individual datasets (*ie* the global model, Table 6.3 Entry 24) is 92.50%. The performance of this global model (Table 6.6, Entry 2-6) is comparable to that of the local models trained from the individual dataset (Table 6.3, Entry 3, 6, 9, 15 and 18) in evaluation with reactions from the individual dataset. The above demonstrates the robustness and generalisability of the global model.

Table 6.6: Performance of '2-bond +' global models when evaluating with individual reaction datasets. The size of the dataset involved in training and testing is specified below in a consistent format. For example, '77(105) first-year' implies 77 sets of first-year reactions with the same reactants, equivalent to 105 reactions.

Ent ry	Testing dataset	% by atoms			% sets of reactants with:		
		Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
Combining all datasets: the global model							
Train: 78 (105) first-year + 100 (205) [3+2] cycloaddition + 100 (100) Diels-Alder + 100 (107) RGD1 + 100 (126) Green,							
Test: 30 (42) first-year + 100 (219) [3+2] cycloaddition + 100 (100) Diels-Alder + 100 (110) RGD1 + 100 (120) Green							
1	All testing data	92.50%	81.08%	75.69%	35.3%	50.7%	54.2%
2	First-year	91.29%	83.02%	68.75%	33.3%	56.7%	53.3%
3	[3+2] cycloaddition	99.14%	96.80%	96.58%	80.0%	94.0%	87.0%
4	Diels-Alder	96.07%	78.84%	83.33%	38.0%	54.0%	65.0%
5	RGD1	77.64%	68.17%	59.03%	6.0%	19.0%	27.0%
6	Green	82.54%	81.11%	70.12%	18.0%	34.0%	38.0%

6.5 Summary

In this investigation, a framework for predicting the site of reactivity has been developed. Machine learning classification models were trained on labels and descriptors at the atomistic level. Descriptor array compositions that cover chemical environments within 'one-bond', 'two-bond' and 'two-bond +' to the atom have been considered.

Thorough evaluations on models based on different descriptor compositions were conducted with various datasets. Improvements in the model performance were observed when the bond strength and neighbours' connectivity components were added to the descriptor array. Decent performance metrics were achieved for the type A (*ie* first-year reaction) and the type B (*ie* the [3+2] cycloaddition and Diels-Alder reaction) datasets, indicating that the chemical environment within two bonds of an atom can determine reactivity to a considerable extent. The applicability of the model framework has been demonstrated with the more complex datasets that cover a diverse chemical space, namely the RGD1 and the Green dataset. Even for these challenging datasets, an accuracy of over 80% has been achieved with a small dataset.

This investigation not only yield a model for predicting reactive sites, but also provide insights into the limits and potentials of data-driven approaches for solving challenges in organic chemistry. Contents of this chapter set a fundament for research into predicting mechanism and potential reaction pathways.

All the scripts and data are available in github.com/Goodman-lab/TwoBondChem.

Recapitulation, Conclusions and Future Works

7.1. Conclusions

The study of organic reactions and their reactivity is a fascinating area of research due to its inherent complexity. Computational analyses provide insights into chemical reactivity and allow us to put a rationale on the experimental outcomes. Using a combination of MM approaches for navigating the conformational space and QM calculations for optimisations of the ground state and TS structures is a well-established approach in investigative reactivity. Such an approach proves to be effective in the study of many organic reactions, including the investigation on 5-methylcytosine pyridination reactions presented in Chapter 2 of this thesis. The reaction mechanism for the pyridination of 5-methylcytosine was elucidated based on kinetics and thermodynamics calculations, which also explain the variations in the percentage yield upon changes in the substituent of the pyridine substrates.

The established workflow with MM and QM simulations is not perfect. Exploring the conformational space of complex and flexible organic molecules poses challenges in terms of computational cost and effectiveness of the calculations. In Chapter 3, an efficient and effective procedure for investigating large and structurally flexible chemical systems was proposed when investigating the reactions with diarylprolinol silyl ether. The origin of enantioselectivity was rationalised from the perspective of conformational changes and structural deformations. Building on top of the computational procedure designed for the diarylprolinol silyl ether catalytic systems, CONFPASS was developed (Chapter 4). CONFPASS is an open-source Python package for fast conformational exploration of organic molecules based on clustering algorithms and machine learning classification models. With CONFPASS, users can finish DFT re-optimisations of conformers from conformational searches at an earlier stage with confidence that the global minimum structure has been obtained.

Sometimes, thermodynamics and kinetics studies alone are not sufficient to explain the outcome of an organic reaction. In Chapter 5, the study on a group of oxazaborolidinium ion-catalysed reactions between aldehyde and diazo compounds shows that transition state theory alone cannot predict the reaction selectivity. The selectivity of the reactions is controlled by reaction dynamics. As this is a large and complex organic system with more than 100 atoms involved, the associated computational cost of running molecular dynamics calculations for confirming the reaction dynamics effect is very costly. Results presented in this thesis have shown that VRAI-selectivity, a Python programme developed within the Goodman group, can predict the chemo- and the stereoselectivity in these reactions. Some general implications on reaction selectivity were concluded from the perspective of the catalyst design and selection of the substrates.

Exploring different reaction pathways and elucidating the reaction mechanism are challenging but unavoidable during computational reactivity investigations. Chapter 6 focuses on understanding the limits and possibilities of data-driven approaches. A method for predicting reaction sites based only on

a simple two-bond model based on an atomistic approach was investigated and evaluated with datasets of various complexity, from simple first-year reactions to the complex RGD1 and Green dataset. There are models which can predict reactions with greater accuracy than this. However, this is a very simple model. Yet, it works well enough to give helpful guidance about reactivity and hints about potential reaction pathways.

Chapter 2: Applying the established computational workflow to understand reactivities in new reactions

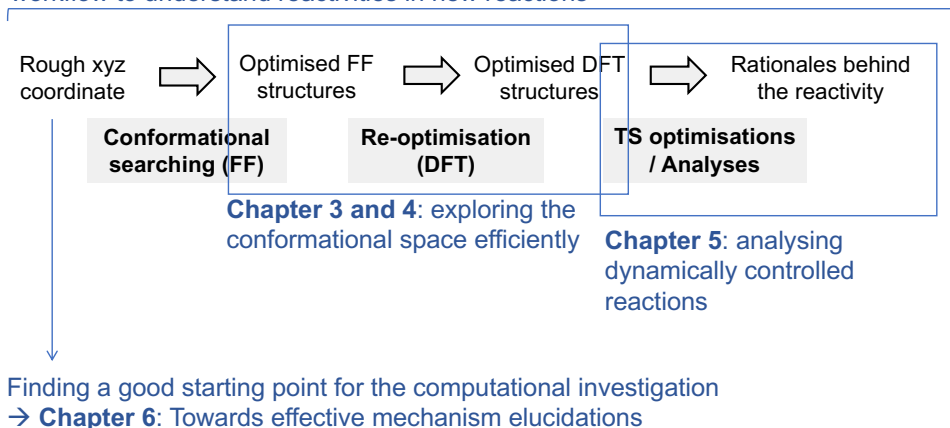


Figure 7.1. Annotations on the computational workflow to understand chemical reactivity (Figure 1.2): contributions to the computational organic chemistry community from this PhD work.

The works from this PhD research provide valuable insights into complex organic reactions, along with practical tools and frameworks for effective and efficient prediction and explanation of reactivity (Figure 7.1).

7.2. Future Works

The work from this thesis can be further developed and improved.

The work on understanding organic reactivity in Chapter 2 and 3 led to the development of CONFPASS in Chapter 4. The applicability domain of CONFPASS can be further extended into inorganic complexes, including organometallics. The transfer learning approach will be explored. A small dataset of inorganic complexes will be generated following the procedure for generating the DFT dataset during the CONFPASS project. The existing model based on organic molecules will be re-purposed and re-used to develop the model for inorganic complexes. Machine learning models for predicting energy of structures at the DFT level, such as MolE8,²⁰⁴ PhysNet²⁰⁶ and ANI potentials²⁰⁷ can be integrated into the CONFPASS pipeline to improve the priority list generating procedure and speed up the process of obtaining low energy conformers.

In Chapter 5, there are still discrepancies between the predicted results with VRAI-selectivity and the experimental results. One potential reason for the discrepancy is that the reaction dynamic control is only partial, while VRAI-selectivity assumes that the selectivity is fully controlled by reaction dynamics. To address this, a possible future work direction is to gather more examples of dynamically controlled reactions to deduce scenarios of partial dynamics controls and improve the VRAI-selectivity algorithm further in coping with such situations.

The model in Chapter 6 predicts atoms involved in connectivity changes during the reaction and so can be used as the starting point in building a pipeline for predicting mechanisms and potential products. The accuracy of the model can also be further improved by considering descriptors that cover the chemical environment beyond two bonds.

As a final remark, the newfound understandings and resources from this thesis provide a solid foundation for further explorations on reactivity and will empower researchers to tackle increasingly complex organic reactions in the future.

Appendix

Notes on Theories

The following sections give a high-level overview of the theory behind DFT calculations. For more details, please refer to the following reference.^{208–210}

DFT calculations provide approximate solutions to Schrodinger's equation. The time-independent Schrodinger's equation is given below, where Ψ is the wavefunction of the system, $\hat{H}$ is the Hamiltonian operator and the eigenvalue, E , is the energy of the system.

$$\hat{H}\Psi = E\Psi \quad (\text{S1})$$

The Hamiltonian operator can be separated into a number of components:

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ne} + \hat{V}_{nn} + \hat{V}_{ee} \quad (\text{S2})$$

$\hat{T}_e$ and $\hat{T}_n$ are operators associated with the kinetic energy of the electrons and nuclei, respectively. $\hat{V}_{ne}$, $\hat{V}_{nn}$ and $\hat{V}_{ee}$ are linked to the potential energy of nucleus-electron, nucleus-nucleus and electron-electron interactions respectively. The Born-Oppenheimer approximation allows us to focus only on the electronic parts ($\hat{T}_e$, $\hat{V}_{ne}$ and $\hat{V}_{ee}$). The approximation recognises that the nuclei in molecules have a much greater mass compared to electrons. Thus, nuclei are assumed to be fixed in position and the motion of electrons can be treated separately.

Hohenberg–Kohn (HK) theorems form the theoretical foundation of DFT. The first HK theorem states that the ground state of a system is uniquely defined by the electron density, $\rho(\mathbf{r})$. $\mathbf{r}$ is the coordinate notation, which includes position and spin coordinates. The total number of electrons, N , can be obtained via the integration of $\rho(\mathbf{r})$ over all space.

$$\int \rho(\mathbf{r}) d\mathbf{r} = N \quad (\text{S3})$$

From the second HK theorem, we can calculate the ground-state energy value, E_0 , of a system based on the vibrational principle when the true density, ρ_0 , is known. The energy value calculated based on trial density, ρ_{trial} , is expected to be greater than the ground-state energy unless the trial density is the same as the true density.

$$E_0 = E_0[\rho_0] \leq E[\rho_{trial}] \quad (\text{S4})$$

In 1965, Kohn and Sham proposed a method to facilitate energy calculation without directly solving Schrodinger's equation. Their proposal involves having a fictitious system of non-interacting electrons, which has the same electron density as for the real system, as the starting point. In this way, the $\hat{H}$ can be treated as the sum of one-electron operators. Sequentially, the energy of the system can be expressed as below:

$$E[\rho(\mathbf{r})] = T_{ni}[\rho(\mathbf{r})] + V_{ne}[\rho(\mathbf{r})] + V_{ee}[\rho(\mathbf{r})] + \Delta T[\rho(\mathbf{r})] + \Delta V_{ee}[\rho(\mathbf{r})] \quad (\text{S5})$$

T_{ni} is the kinetic energy of the non-interacting electrons, while V_{ne} and V_{ee} are the classical electron-nucleus and electron-electron interaction energy. These terms are relatively straightforward to define.

ΔT and ΔV_{ee} are the correction terms for accounting non-classical electronic interactions. These two terms form the exchange-correlation energy, E_{xc} . The exact form of exchange-correlation functionals is not known. Approximations need to be made and this is an important field of research within the computational chemistry community. The DFT functionals are often named according to the approach used to model the exchange-correlation function.

There are different levels of sophistication in approximating E_{xc} . This is described by Jacob's ladder (Figure S1.1). The simplest function is based on the local density approximation (LDA). Here, the electron density is assumed to be uniform across space and the value of E_{xc} is calculated exclusively based on the local electron density. Alternatively, the accuracy of the functional can be improved by also considering the gradient of the electron density, which provides information on changes in electron density over space. This approach is called the generalised gradient approximation (GGA). Moving up the ladder, the meta-GGA approach includes the Laplacian of the density or the kinetic energy density on top of just the density and gradient of the density. The hybrid approach can be adopted to improve the accuracy further. GGA approaches tend to systematically underestimate energy barriers, while the Hartree-Fock (HF) approach overestimates energy due to the lack of descriptions of correlation effects. In the hybrid functionals, the HF exchange term is mixed with the DFT exchange-correlation functional to give improved energy estimation accuracy. The random phase approximation (RPA) and doubly hybrid density functionals are nearly at the top of Jacob's ladder. Whilst the improved accuracy, these methods are very computationally expensive.

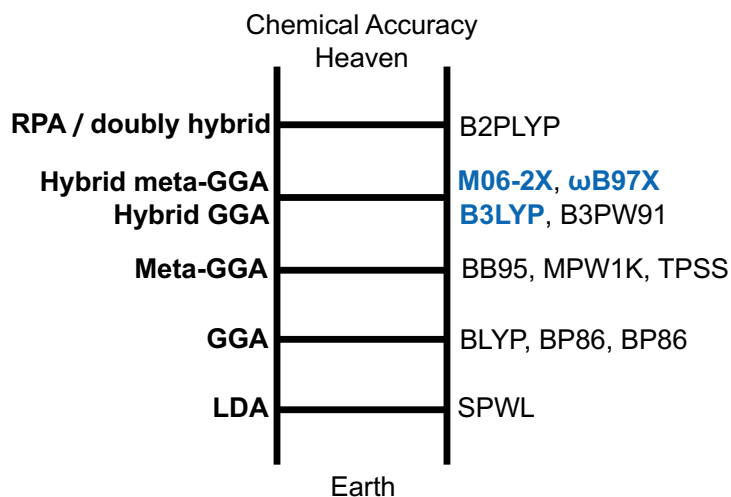


Figure S1.1. Jacob's ladder on the level of theory in DFT calculations.^{211,212} The most used functionals in this thesis (B3LYP, ω B97X and M06-2X) are coloured in blue.

B3LYP,^{105–107} ω B97X¹⁰⁹ and M06-2X²¹³ are the most used functionals in this thesis. They are chosen for a balance between good accuracy and the amount of required computational resources. B3LYP is a 3-parameter hybrid GGA functional. B3LYP is a modification of the B3PW91 functional from Beck, where the correlation function from Lee, Yang and Parr (LYP) is used in instead of PW91 functional.

ω B97X is a hybrid meta-GGA functional developed by Head-Gordon *et al.*¹⁰⁹ M06-2X belongs to the Minnesota family by Truhlar *et al.* and is also a hybrid meta-GGA functional.²¹³

In this work, B3LYP and ω B97X are used with empirical dispersion correction terms from Grimme *et al.*²⁰⁹ as B3LYP-D3 and ω B97XD. The ω B97XD functional in Gaussian 16 uses the D2 dispersion model.

The DFT energy with dispersion corrections, E_{DFT-D} , is given as:

$$E_{DFT-D} = E_{DFT} + E_{disp} \quad (S6)$$

E_{DFT} is the DFT energy from using the chosen functional and E_{disp} is the empirical dispersion correction energy. The D2 model²¹⁴ accounts for two-body dipole-dipole dispersion effects.

$$E_{disp}^{D2} = -s_6 \sum_{AB} \frac{c_6^{AB}}{R^6} f_{damp}^{Fermi}(R) \quad (S7)$$

s_6 is a scaling factor and depends on the chosen DFT functional. c_6^{AB} , the dispersion coefficient for the atom pair A and B, includes all components from dipole-dipole interactions. R is the atomic distance between A and B. A dumping function is required to decrease the dispersion energy at short interatomic distance. Here, a Fermi-type damping function, $f_{damp}^{Fermi}(R)$, has been chosen. The dispersion correction energy in D3 dispersion model²¹⁵ consists of two components:

$$E_{disp}^{D3} = E_{disp}^{2-body,D3} + E_{disp}^{ATM,D3} \quad (S8)$$

$E_{disp}^{2-body,D3}$ is the total atom pairwise energy.

$$E_{disp}^{2-body,D3} = - \sum_{AB} \sum_{n=6,8} s_n \frac{c_n^{AB}}{R^n} f_{damp}^{(n)}(R) \quad (S9)$$

Here, the higher-order multipole term c_8^{AB} is included to account for dipole-quadrupole interactions. $E_{disp}^{ATM,D3}$ is the dipole-dipole-dipole dispersion energy or the Axilrod-Teller-Muto term, ATM.

$$E_{disp}^{ATM,D3} = - \sum_{ABC} \frac{c_9^{ABC} (3\cos\theta_a \cos\theta_b \cos\theta_c + 1)}{(R_{AB} R_{AC} R_{CB})^3} f_{damp}^{(9)}(\overline{R_{ABC}}) \quad (S10)$$

$E_{disp}^{ATM,D3}$ describes dispersions in three-body systems. θ_a , θ_b and θ_c are the internal angles of the triangle formed by atom A, B and C. $\overline{R_{ABC}}$ is the mean of the interatomic distance R_{AB} , R_{AC} and R_{CB} . The D3 model uses the damping model from Head-Gordon *et al.*¹⁰⁹ The D2 and D3 dispersion correction model are very computationally efficient as only the geometry is required as input.

A basis set is required as input for DFT calculations. A basis set consists of mathematical functions that construct electronic wavefunctions. Calculations in this thesis employed the split-valence basis sets from Pople *et al.*²¹⁰ 'Split-valence' implies that core orbitals are described by a single basis function, whereas valence orbitals are represented by multiple basis functions. In the Pople basis set, the notation follows the format X-YZG when the valence orbitals are described by two sets of basis functions, *ie* double-zeta. Examples include 3-21G, 4-21G and 6-31G. The first number indicates the number of primitive Gaussian functions used for the core orbitals. The values after '-' represent the number of primitive Gaussian functions in linear combination in each set of the valence basis functions. Triple-zeta basis sets, such as 6-311G, are also possible and were used in thesis. Polarisation functions can be added to the basis sets, *ie* as (d,p) at the end of the notation. (d) represents adding a set of d functions

to polarise the p functions. (p) implies that p functions are added to H and He. The addition of diffuse functions is often required to describe excited electron states and high-energy molecular orbitals accurately. Diffuse functions are indicated by '+' or '++' symbols in the basis set notation. '+' means that all additional sets of s and p functions with small exponents have been added to the heavy atoms. On top of this, diffuse s functions can also be included in the hydrogen atoms, and this is marked by the '++' notation. In the development of CONFPASS (Chapter 4), DFT data from Grayson *et al.*⁵² was also used. Some of their calculations were carried out with basis set from Ahlrichs *et al.*,²¹⁶ specifically def2-TZVPP. def2-TZVPP is a triple-zeta basis set with additional polarisation functions.

Notes on calculations

Unless expressly stated, the following settings were used:

Conformational searching calculations with MacroModel

A setting of mixed torsional/low-mode sampling method and a maximum number of steps per rotatable bond was used. The default setting (*ie* 0.5 Å) was used for the RMSD cut-off to filter conformers with similar structures.

Data analyses and diagrams

Data analyses, such as processing outputs from DFT calculations, were conducted with Python (3.8.12). Graphs were created with Matplotlib (3.3.2) or Plotly (5.1.0).²¹⁷ 3D images of the optimised structures were generated with CYLview20²¹⁸, PyMOL²¹⁹ and Visual Molecular Dynamics (VMD)²²⁰.

Density functional theory

Density functional theory (DFT) calculations were performed with Gaussian 16.¹⁰⁴ Optimised structures were verified through frequency analyses. All the geometries were confirmed to correspond to a minimum or a first-order saddle point on the potential surface. Intrinsic reaction coordinate (IRC) or quick reaction coordinate (QRC) calculations¹⁴¹ were conducted to further verify that the identified TSs correspond to the relevant processes of interest.

Appendix to Chapter 2

2.A Data availability

The structural information is available in the Cambridge Apollo Repository:

<https://doi.org/10.17863/CAM.96455>

The key structures are included in the 'Key_conformation_SI_v10122022.zip' file as opt+freq Gaussian calculation output files.

2.B Overall energy diagram

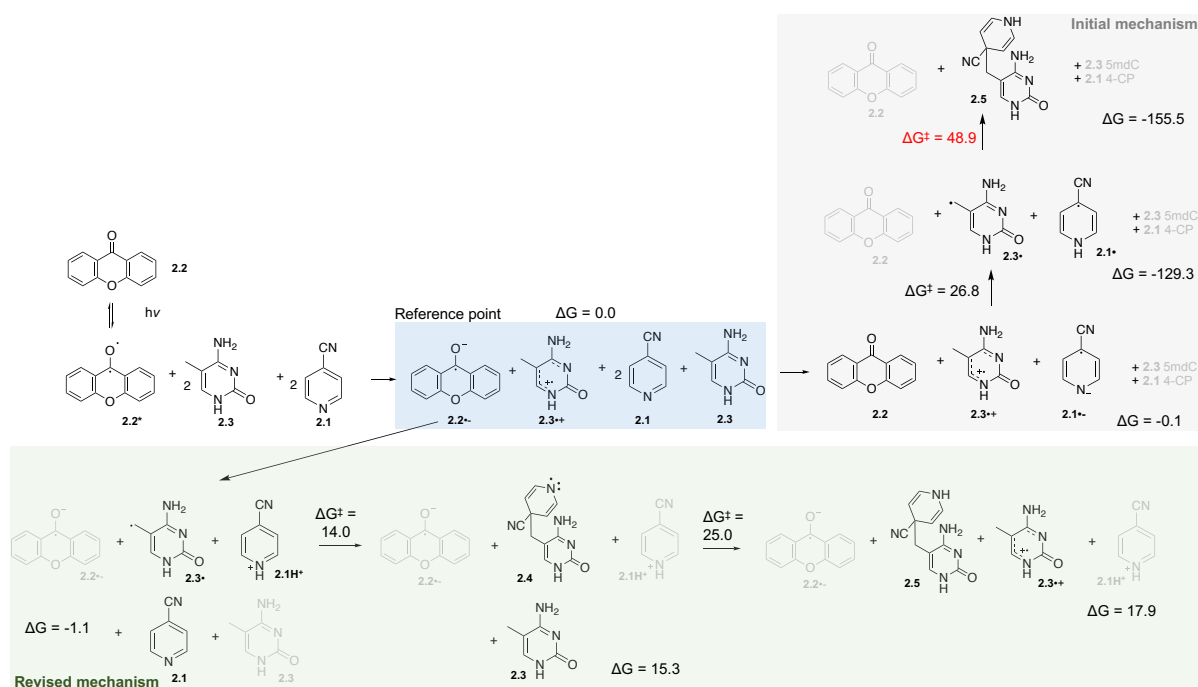


Figure S2.1. An overall energy diagram comparing the initial and revised mechanism. For clarity, structures that stay the same before and after a step in the process are coloured in grey. The radical-radical coupling pathway in the initial mechanism required a high kinetic barrier. Experimentally, the chance of a collision between two radicals is also low as a catalytic amount of xanthone was used. The calculations support the kinetic feasibility of the revised mechanism compared to the original proposal.

Appendix to Chapter 3

3.A Data availability

The structural information is available in the Cambridge Apollo Repository:

<https://doi.org/10.17863/CAM.84437>

The key structures are included in the 'SI_key_structure.zip' file as Gaussian calculation output files. All the structures are the lowest energy conformer in their corresponding group. Both the opt+freq calculation and its associated single point energy output (*ie* _solv_wb.out files) are included for each structure. The corresponding level of theory is ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d).

The relevant scripts for facilitating the pipeline have been compiled into several Jupyter notebooks (.ipynb files) under the Script_im directory in GitHub (https://github.com/Goodman-lab/Script_im). The summary of the files, instructions and examples are given in the notebooks and the README.md file.

```
Script_im/
├── B_Me_RMe_R.xyz
├── C_Me_RMe_dft_r2.csv
├── C_Me_RMe_R.xyz
├── conformation_label_generation.ipynb
├── Filter_diversity_incl_label_generation_and_selection.ipynb
├── gjf_generation.ipynb
├── Goodvibes_output_B_Me_RMe_R_TS.xyz
├── Goodvibes_output_C_Me_RMe_R_TS.xyz
├── Goodvibes_output_test.xyz
├── README.md
├── ring-out.csv
└── Selection_remove_repetitive_structures.ipynb
```

3.B Dihedral angle distribution studies

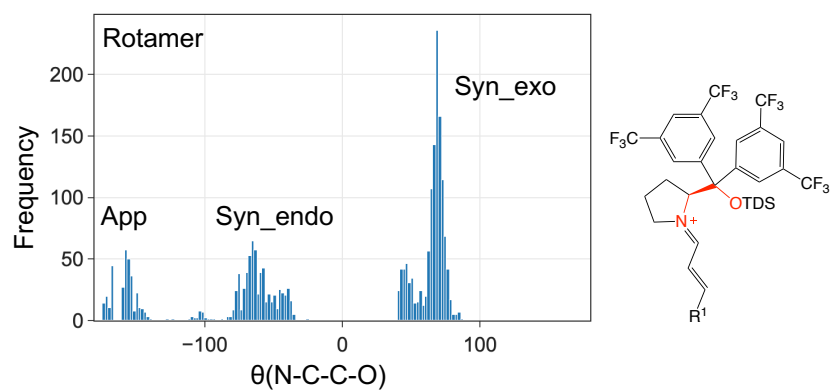


Figure S3.1. The dihedral angle distribution for the 'rotamer' structural feature.

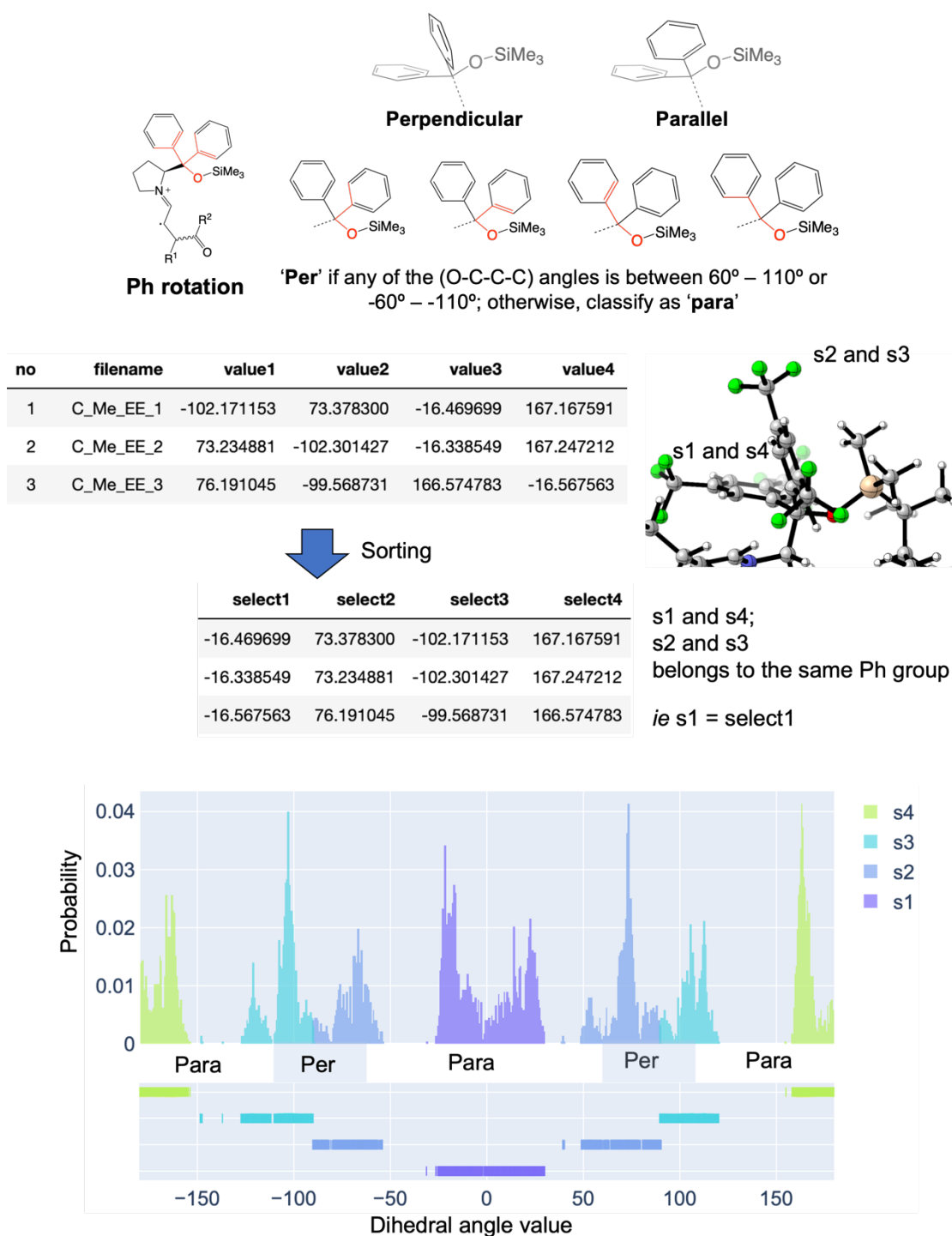


Figure S3.2. The dihedral angle distribution for the 'Ph rotation' structural feature. The probability vs dihedral angle value plot is presented above. The structures are at the MMFF level from the conformational searching calculation of catalyst C iminium. The dihedral value data was sorted before the graph plotting. Cut-off values that minimise the number of borderline examples were chosen. 'Per' and 'para' are labels for different Ph rotation orientation. 'Per' means that the Ph rings are perpendicular to each other. 'Para' means that the Ph rings are parallel to each other.

3.C Benchmarking

Energetic and structural datasets at the ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d) level of theory are as below:

Table S3.1. The computational kinetic data at the ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d) level of theory (Unit: kcal mol⁻¹)

	R¹ = Me; R² = Me		R¹ = Me; R² = Ph		R¹ = Ph; R² = Me		R¹ = Ph; R² = Ph	
	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$
Catalyst A								
SR	10.3	0.0	12.2	0.4	-	-	-	-
SS	11.2	0.9	11.7	0.0	-	-	-	-
Catalyst A_OTDS								
SR	9.8	0.0	11.6	0.0	-	-	-	-
SS	11.2	1.4	12.0	0.4	-	-	-	-
Catalyst B								
SR	10.2	0.0	11.8	0.3	9.4	0.0	12.4	0.0
SS	10.9	0.7	11.5	0.0	12.6	3.2	14.7	2.3
Catalyst C								
SR	10.5	0.0	11.6	0.0	-	-	11.2	0.0
SS	12.7	2.2	13.6	2.0	-	-	13.7	2.5

Table S3.2. The conformation label of the lowest energy TSs at the ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d) level of theory. All the structures presented below have a 'trans-ene-iminium' label.

R ¹ = Me; R ² = Me		R ¹ = Me; R ² = Ph		R ¹ = Ph; R ² = Me		R ¹ = Ph; R ² = Ph	
Catalyst A							
SR	syn_exo EE 3 out per	syn_exo E 3 out per		-		-	
SS	syn_exo EE 3 in per	syn_exo E 3 in per		-		-	
Catalyst A_OTDS							
SR	syn_exo EE 3 out per 1	syn_exo EE 3 out per 1		-		-	
SS	syn_exo EE 3 in per 0	syn_exo EE 3 in per 0		-		-	
Catalyst B							
SR	syn_exo EE 3 out per	syn_exo EE 3 out per	syn_exo EE 4 out per		syn_exo EE 3 out per		
SS	syn_exo EZ 3 out per	syn_exo EZ 3 out per	syn_exo EE 3 in per		syn_exo EZ 3 out per		
Catalyst C							
SR	syn_exo EE 4 out per 1	syn_exo EE 4 out per 1		-		syn_exo EE 0 out per 1	
SS	syn_exo EE 4 in per 1	syn_exo EZ 4 in per 1		-		syn_exo EZ 5 out per 1	

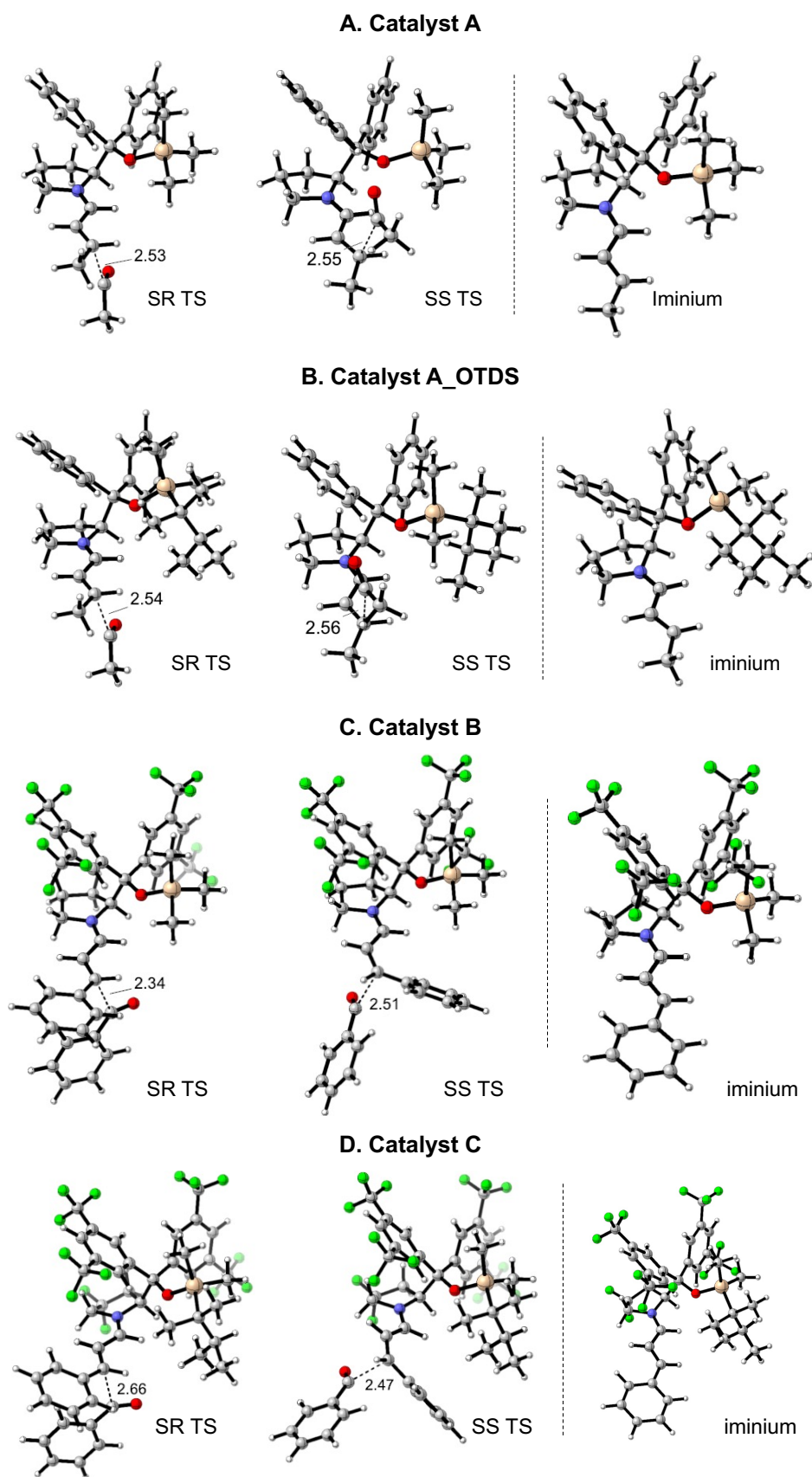


Figure S3.3. The structure of the lowest energy TS and iminium structures ($R^1 = R^2 = \text{Ph}$). A) Catalyst A, B) Catalyst A_OTDS, C) Catalyst B and D) Catalyst C systems. The bond length values of the forming C-C bonds are labelled in the diagrams.

Table S3.3. The computational kinetic data of the lowest energy TSs at the ω B97X-D/6-311++G(d,p)//B3LYP/6-31G(d) level of theory. (Unit: kcal mol⁻¹)

	R¹ = Me; R² = Me		R¹ = Me; R² = Ph		R¹ = Ph; R² = Me		R¹ = Ph; R² = Ph	
	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$	$\Delta G^\ddagger$	$\Delta\Delta G^\ddagger$
Catalyst A								
SR	7.3	0.0	6.7	0.5	-	-	-	-
SS	8.2	0.9	6.2	0.0	-	-	-	-
Catalyst A_OTDS								
SR	6.3	0.0	5.6	0.0	4.6	0.0	-	-
SS	7.9	1.7	6.3	0.7	6.5	1.9	-	-
Catalyst B								
SR	5.6	0.0	4.5	0.0	5.2	0.0	4.6	0.0
SS	6.6	1.1	4.6	0.1	6.5	1.3	5.6	1.0
Catalyst C								
SR	4.8	0.0	3.3	0.0	-	-	2.6	0.0
SS	7.9	3.1	4.3	1.0	-	-	3.7	1.1

Table S3.4. The conformation label of the lowest energy TSs at the ω B97X-D/6-311++G(d,p)//B3LYP/6-31G(d) level of theory. All the structures presented below have a 'trans-ene-iminium' label. (Unit: kcal mol⁻¹)

	R ¹ = Me; R ² = Me	R ¹ = Me; R ² = Ph	R ¹ = Ph; R ² = Me	R ¹ = Ph; R ² = Ph
Catalyst A				
SR	syn_exo EE 3 out per	syn_exo EE 3 out per	-	-
SS	syn_exo EE 3 in per	syn_exo EE 3 in per	-	-
Catalyst A_OTDS				
SR	syn_exo EE 3 out per 1	syn_exo EE 3 out per 1	syn_exo EE 3 out per 1	-
SS	syn_exo EE 3 in per 0	syn_exo EE 3 in per 0	syn_exo EE 3 in per 0	-
Catalyst B				
SR	syn_exo EE 3 out per	syn_exo EE 3 out per	syn_exo EE 4 out per	syn_exo EE 3 out per
SS	syn_exo EE 3 in per	syn_exo EZ 3 out per	syn_exo EE 3 in per	app EE 3 in para
Catalyst C				
SR	syn_exo EE 4 out per 1	syn_exo EE 4 out per 1	-	syn_exo EE 0 out per 1
SS	app EE 4 in per 1	app EE 4 in per 1	-	app EE 4 in per 1

3.D Distortion-interaction analyses

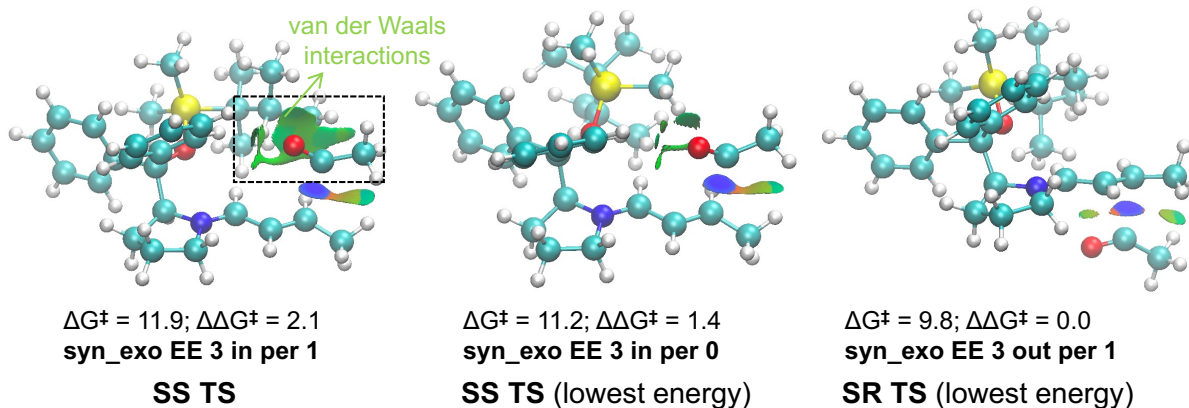
Table S3.5. Distortion/interaction analyses. The conformation label and $\Delta\Delta G^\ddagger$ of the TSs are given along with $\Delta U_{d,iminium}^\ddagger$ from distortion/interaction analyses. The kinetic ($\Delta\Delta G^\ddagger$) and structural data are given at the $\omega B97X-D/6-311++G(d,p)/SMD/CH_3CN//B3LYP/6-31G(d)$ level of theory. All the TSs presented below have a 'trans-ene-iminium' label. (Unit: kcal mol⁻¹)

Entry	Catalyst	SR/SS?	$\Delta\Delta G^\ddagger$ (TS)	$\Delta U_{d,iminium}^\ddagger$	Conformation label (TS)
R¹ = Me; R² = Me					
1	A	SR	0.0	2.4	syn_exo E 3 out per
2	A	SS	0.9	3.4	syn_exo E 3 in per
3	A_OTDS	SR	0.0	2.5	syn_exo EE 3 out per 1
4	A_OTDS	SS	1.4	3.6	syn_exo EE 3 in per 0
5	A_OTDS	SS*	2.0	4.6	syn_exo EE 3 in per 1
6	B	SR	0.0	2.1	syn_exo EE 3 out per
7	B	SS	0.7	2.5	syn_exo EE 3 in per
8	C	SR	0.0	1.1	syn_exo EE 4 out per 1
9	C	SS	2.2	3.6	syn_exo EE 4 in per 1
R¹ = Me; R² = Ph					
10	A	SR	0.4	4.0	syn_exo E 3 out per
11	A	SS	0.0	4.1	syn_exo E 3 in per
12	A_OTDS	SR	0.0	4.0	syn_exo EE 3 out per 1
13	A_OTDS	SS	0.4	4.2	syn_exo EE 3 in per 0
14	A_OTDS	SS*	2.5	7.5	syn_exo EE 3 in per 1
15	B	SR	0.5	4.0	syn_exo EE 3 out per
16	B	SS	0.0	1.7	syn_exo EZ 3 out per
17	B	SS*	5.1	5.9	syn_exo EE 3 in per
18	C	SR	0.0	4.4	syn_exo EE 4 out per 1
19	C	SS	2.0	6.2	syn_exo EE 4 in per 1
R¹ = Ph; R² = Ph					
20	B	SR	0.0	4.6	syn_exo EE 3 out per
21	B	SS	2.3	2.0	syn_exo EZ 3 out per
22	B	SS*	5.4	10.1	syn_exo EE 4 in per
23	C	SR	0.0	3.4	syn_exo EE 0 out per 1
24	C	SS	2.5	2.8	syn_exo EZ 5 out per 1
25	C	SS*	5.7	12.8	syn_exo EE 4 in per 1

* The structure is not the lowest energy conformer of its group.

3.E Non-covalent interaction analyses

Catalyst A_OTDS; R¹ = R² = Me systems



Catalyst B; R¹ = R² = Ph systems

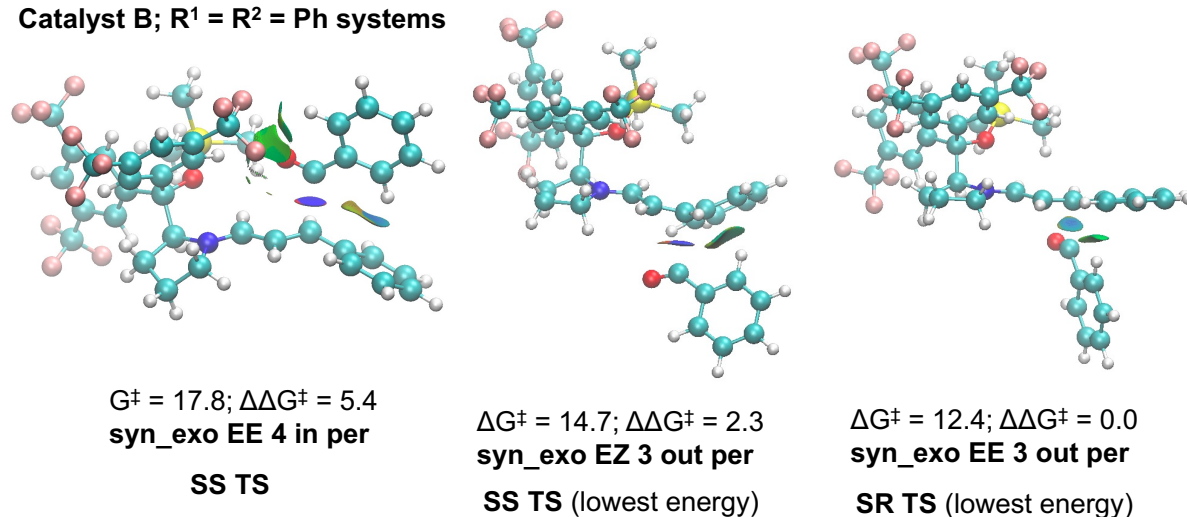


Figure S3.4. Non-covalent interactions (NCIs) analyses on the TSs. The colour of the surfaces within the TS structures indicates the nature of the interactions (*ie* blue = attractive, red = repulsive and green = weak van der Waals interactions). The lowest energy TSs with the specific stereochemistry are labelled in the diagram.

(Level of theory: ω B97X-D/6-311++G(d,p)/SMD/CH₃CN//B3LYP/6-31G(d), unit: kcal mol⁻¹)

3.F Overall energy diagram

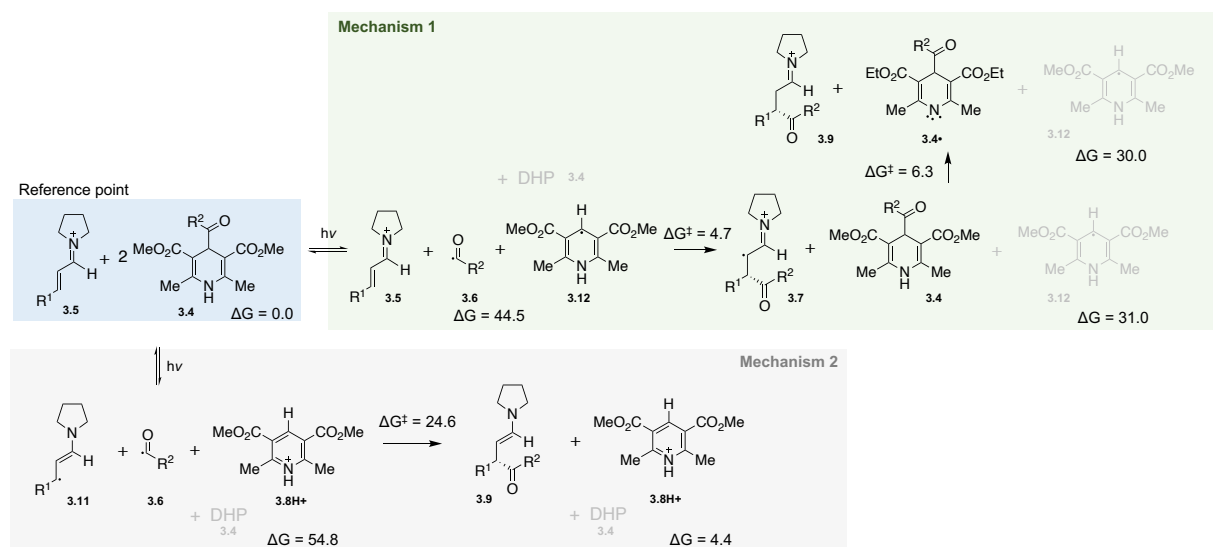


Figure S3.5. An overall energy diagram comparing mechanism 1 and 2. For clarity, structures that stay the same before and after a step in the process are coloured in grey.

Appendix to Chapter 4

4.A Data and scripts

CONFPASS stands for '**Conformer Prioritisations & Analysis** for DFT re-optimisations' (available to download via: github.com/Goodman-lab/CONFPASS). The package is written entirely in Python and developed under the following environment:

- The Python (3.8.12) Standard Library (os, sys, itertools, collections, pickle, random, optparse, traceback, json)
- Pandas (1.3.4)
- Numpy (1.21.2)
- Sklearn (1.0.1)¹³²
- RDkit (2019.09.3)⁴²
- Natsort (8.0.2)

The script from <https://github.com/jensengroup/xyz2mol/blob/master/xyz2mol.py> (which is based on the work of Kim *et al.*, last accessed May 16, 2024)¹⁵² was used to convert xyz coordinates to a RDkit mol object.

This package is suitable for processing organic molecules, including radical species, and is currently under development for processing organometallics and inorganic compounds. More details can be found in the user guide, where examples are given on the execution of the package as a Python module. The package can also be imported and used in python scripts. Documentation and tutorials are provided in a Jupyter notebook. Example SDFs from conformational searching calculations and output file folders are also given.

The directory tree of the CONFPASS package (v19042023):

```
confpass/  
├── __init__.py  
├── __main__.py  
├── BinsTest_rmsCheck.py  
├── cal_dihedral_v2.py  
├── clustering_dih_v7.py  
├── confpass.py  
├── correcting_dihedral_v1.py  
├── dihedral_parameter_v2.py  
├── get_dft_output_v2.py  
├── GetPriority_v3.py  
├── isolate_key_dihedral_v5.py  
├── LR_model_24042023_x10.sav  
├── MolFrac_ML.py  
├── to_g16_input_v2.py  
└── xyz2mol.py
```

The directory tree of the user guides, SDFs and output file folders:

```
test_demo/
├─ CONFPASS_tutorial_v4.ipynb
├─ omegacsd_VORJER.sdf
├─ test_15.sdf
├─ test_15/
│   └─ ... (Gaussian .out files)
│       └─ spe/
│           └─ ... (Gaussian _spe.out files)
├─ radical/
│   └─ radical_gu_liu1/
│       └─ ... (Gaussian .out files)
│           └─ spe/
│               └─ ... (Gaussian _spe.out files)
└─ radical_gu_liu1.sdf
```

In the DFT dataset, there are three radical molecules. The structure of the radicals and pseudo structures for conformational searching are given in Figure S4.1.

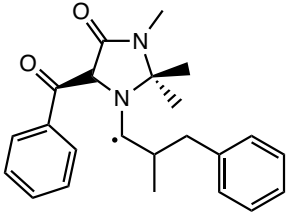
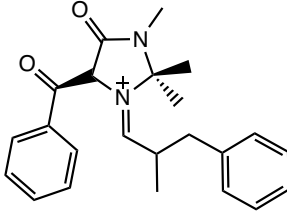
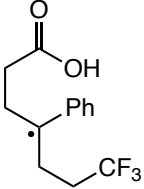
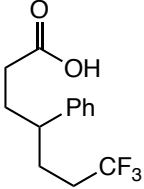
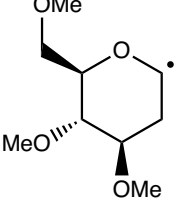
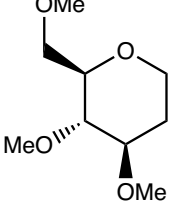
Name	Structure	Pseudo structure
radical_wong1 ¹⁴⁹		
radical_gu_liu1 ¹⁵⁰		
radical_baran3 ¹⁵¹		

Figure S4.1. The structure and pseudo structure of the radical molecules in the DFT dataset. The pseudo structures were used for the conformational searching as MMFF cannot cope with radical species.

4.B Generating CONFPASS priority lists

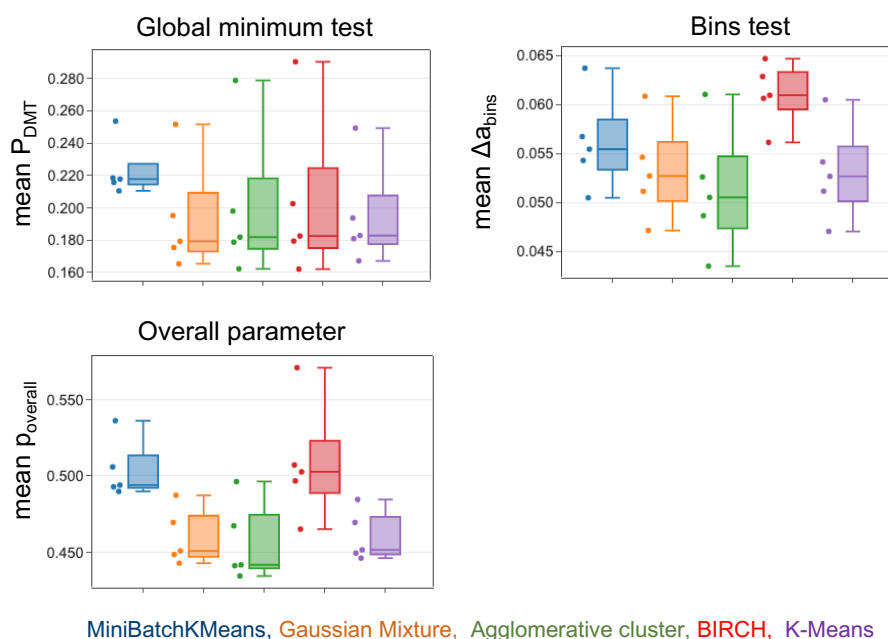


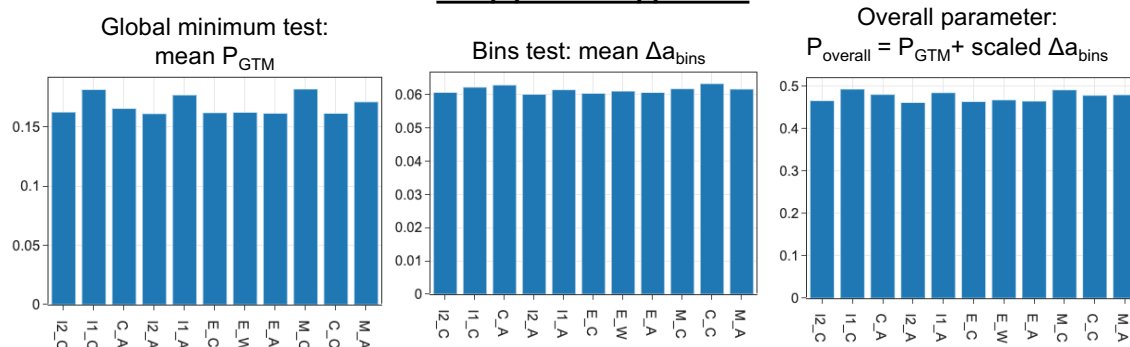
Figure S4.2. The benchmarking test on clustering algorithms: box plot analyses on performance metrics. The box plots in the graphs are colour-coded according to the clustering methods. The clustering algorithms were tested in combination with various priority list generation methods, including *pipeline-x*, *pipeline-ascent*, and *pipeline-mix* ($Q = 0.15, 0.20$ and 0.25 ; $x = 0.8$). Hence, there are 6 data points in each box plot.

Affinity (Metric used to calculate the linkage)

Linkage (The linkage criterion determines which distance to use between sets of observation)

	Ward	average	complete
Euclidean	E_W	E_A	E_C
I1	NA	I1_A	I1_C
I2	NA	I2_A	I2_C
Manhattan	NA	M_A	M_C
cosine	NA	C_A	C_C

The pipeline-x approach



The pipeline-ascent approach

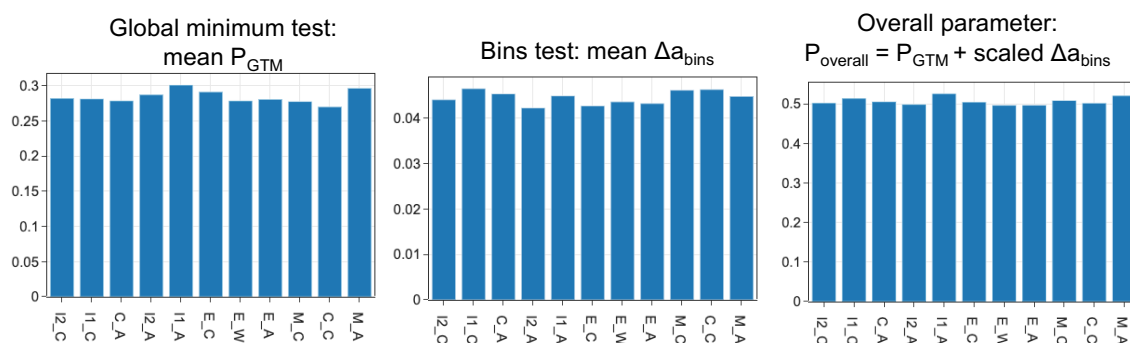


Figure S4.3. Hyperparameter tuning tests for the agglomerative clustering method. Abbreviations are given for combinations of linkage and affinity settings as shown in the table. The Ward linkage method only works with Euclidean affinity setting.

4.C Post-processing procedures

Introducing post-processing procedures only leads to marginal improvements in the global minimum test measurement, mean P_{GMT} . Variations in bins test performance mean Δa_{bins} have been observed, but none of the approaches with post-processing steps outperforms the *pipeline-ascent* approach.

Table S4.1. The performance metrics for the clustering approach *pipeline* + post-processing process approach. The overall parameter (P_{overall}), global minimum test measurement (P_{GMT}) and bins test parameter (Δa_{bins}) data are presented below. The data for the *pipeline-x*, *pipeline-ascent* and *ascent* approach are also included for comparisons.

Approach	P_{overall}	P_{GMT}	Δa_{bins}
rv1B3_1	0.461	0.156	0.061
rv1B3_5	0.461	0.156	0.061
rv5B2_1	0.445	0.157	0.058
rv5B2_5	0.459	0.161	0.060
rv1B_1	0.464	0.156	0.061
<i>Pipeline-ascent</i>	0.495	0.276	0.044
<i>Pipeline-x</i>	0.468	0.162	0.061
<i>ascent</i>	0.590	0.165	0.085

Table S4.1 presents the top 5 best-performed clustering approaches with post-processing steps ranked by the overall parameter. Brief descriptions of the post-processing steps in these clustering approach pipelines are given below:

rv1B3_1: The starting point was a priority list generated via the *pipeline-x* approach with $x = 0.8$. When the percentage of re-optimised conformers = 20%, a new priority list was generated via the *pipeline-x* approach with the conformational search output SDF. The percentage of duplicate conformers in the group of re-optimised conformers was deduced via RMSD calculations. If the percentage of duplicated conformers is greater than 20%, x is then set to $(1 - \text{percentage of duplicated conformers})$. The $x = 0.8$ setting was used if the percentage of duplicate conformers was less than 20%. The re-optimisation process continues with the new priority list. Conformers that have already been re-optimised are ignored in the new priority list.

rv1B3_5: The process in rv1B3_1 was repeated for 4 times when the percentage of re-optimised conformers = 20%, 40%, 60% and 80% respectively.

rv5B2_1: The procedure in rv1B3_1 was performed. Additionally, ΔG values of the re-optimised conformers were calculated. The list of conformer clusters from the clustering process was sorted according to the clusters-distance array as shown in Figure S4.4. The new priority list is constructed from the sorted conformer clusters list.

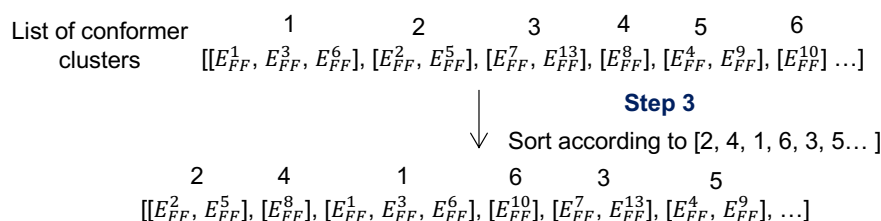
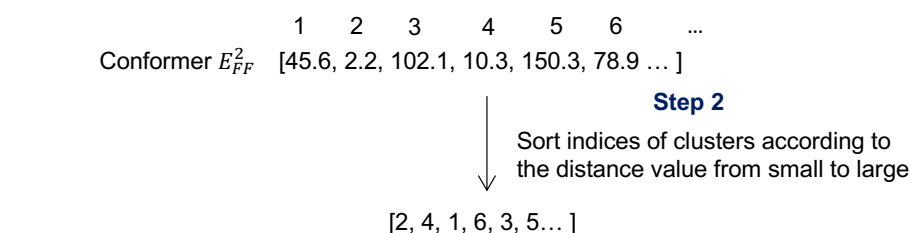
Step 1: get cluster-distance array

The distance to cluster:

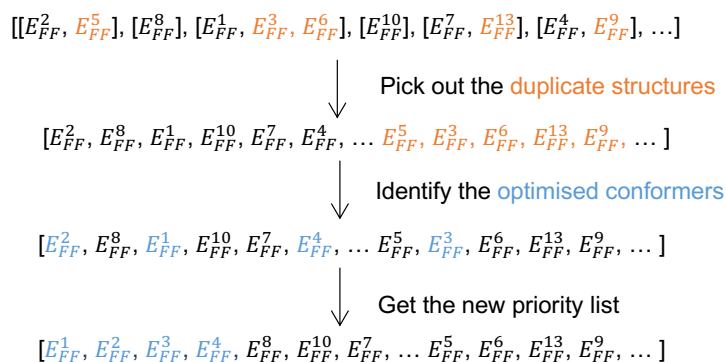
	1	2	3	4	5	6	...
Conformer E_{FF}^1	74.3	57.2	12.6	422.1	230.1	54.9	...
Conformer E_{FF}^2	45.6	2.2	102.1	10.3	150.3	78.9	...
Conformer E_{FF}^3	...						
...							

$M \times N$ array (M number of clusters x N number of conformers)

The re-optimised conformer with the lowest $\Delta G(\text{DFT})$: Conformer 2



Step 4: derive the priority list



Re-optimised conformers are sorted using the old priority list

Figure S4.4. Using the cluster-distance array to sort a list of conformer clusters and derive the new priority list. A 4-step process is followed. The conformer indexes are colour-coded to illustrate the sorting process.

rv5B2_5: The process in rv5B2_1 was repeated for 4 times when the percentage of re-optimised conformers = 20%, 40%, 60% and 80% respectively.

rv1B_1: The starting point was a priority list generated via the *pipeline-x* approach with $x = 0.8$. The output from the clustering process for constructing the priority list (*ie* the list of conformer clusters when $n_clusters = 0.8 \times \text{total number of conformers}$) was retained. When the percentage of re-optimised conformers = 20%, ΔG values of the re-optimised conformers are calculated and a new priority list was constructed following the sorting process in Figure S4.5.

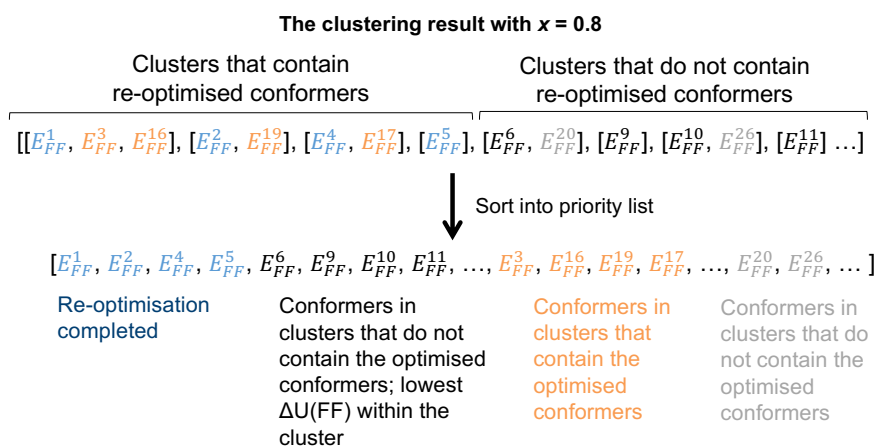


Figure S4.5. The priority list derivation method in rv1B_1. The conformer indexes are colour-coded to illustrate the sorting process.

4.D Evaluating CONFPASS priority lists

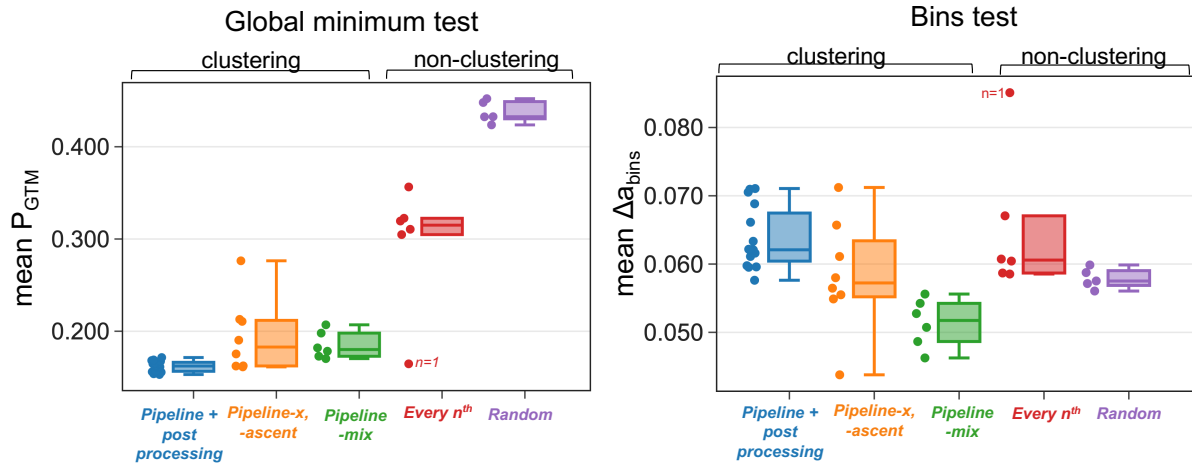


Figure S4.6. Comparisons between different clustering and non-clustering priority list generation approaches. The global minimum test (P_{GTM}) and bins test (Δa_{bins}) results are presented in the box plots. This figure can be cross-referenced with Figure 4.19 in the main text. The tests for the random approach have been repeated 5 times, which contributes to the 5 data points in the category. For *every n^{th} approach*, the tests have been repeated at $n = \{1, 2, \dots, 6\}$, which leads to the 6 data points in the category.

4.E Benchmarking: different levels of theory

Here is a summary of the calculations and the associated levels of theory:

- MMFF $\rightarrow$ ω b97xd/6-311g(d,p)//B3LYP-D3/6-31g(d): the 150 molecules
- MMFF $\rightarrow$ ω b97xd/6-311++g(d,p)// ω b97xd/6-31g(d): the 20 molecules from the Grayson dataset, which are part of the 150 molecules (*ie* the Grayson molecules)

The existing dataset from Grayson *et al.*⁵² were processed and formatted for benchmarking at different levels of theory:

- MMFF $\rightarrow$ M06-2X/def2-TZVPP/IEFPCM/benzene//M06-2X/6-31G(d)/IEFPCM/benzene:
20 Grayson molecules
- MM2 $\rightarrow$ M06-2X/def2-TZVPP/IEFPCM/benzene//M06-2X/6-31G(d)/IEFPCM/benzene:
17 Grayson molecules
- OPLS3e $\rightarrow$ M06-2X/def2-TZVPP/IEFPCM/benzene//M06-2X/6-31G(d)/IEFPCM/benzene:
20 Grayson molecules

The CONFPASS workflow was also evaluated by considering the CREST conformational searching programme:

- GFN2-xTB $\rightarrow$ ω b97xd/6-311++g(d,p)// ω b97xd/6-31g(d): the 20 randomly selected molecules from the master dataset

The datasets of the 20 Grayson molecules at different levels of theory are gathered or generated to test the generality of the clustering approach pipeline. The global minimum test and bins test were performed and the overall parameters were calculated for these DFT datasets (Figure S4.7).

The 'MMFF $\rightarrow$ ω B97XD' and 'OPLS3e $\rightarrow$ M06-2X' datasets produce similar results (*ie* in terms of the overall parameter and mean P_{GMT} vs mean Δ_{abins} plot pattern) as extensive DFT dataset of 150 molecules at the ω B97XD/6-311G(d,p)//B3LYP-D3/6-31G(d) level of theory from MMFF structures. The 'MM2 $\rightarrow$ M06-2X' and 'MMFF $\rightarrow$ M06-2X' dataset shows greater deviations, which may be explained by two reasons. Firstly, the 20 Grayson molecules are more rigid with fewer conformers and more aromatic atoms than other molecules in the DFT dataset on average (Figure S4.8). Secondly, more molecules need to be considered, *ie* a dataset of 20 molecules is insufficient to reach convergence. The below procedures were followed to test the hypothesis:

1. Randomly select X% of molecules from the DFT dataset (total of 150 molecules) and their priority list for evaluation
2. Perform the tests (the bins test and global minimum test)
3. Repeat the above step 1-2 10 times
4. Calculate the mean and standard deviation for the parameters from the tests

The above procedure has been repeated for X% between 5% to 50%. The result is presented in Figure S4.9. as the mean and standard deviation of the overall parameter value vs X% plots. At least 30% of the DFT dataset or an equivalent of 45 molecules are required to achieve convergence.

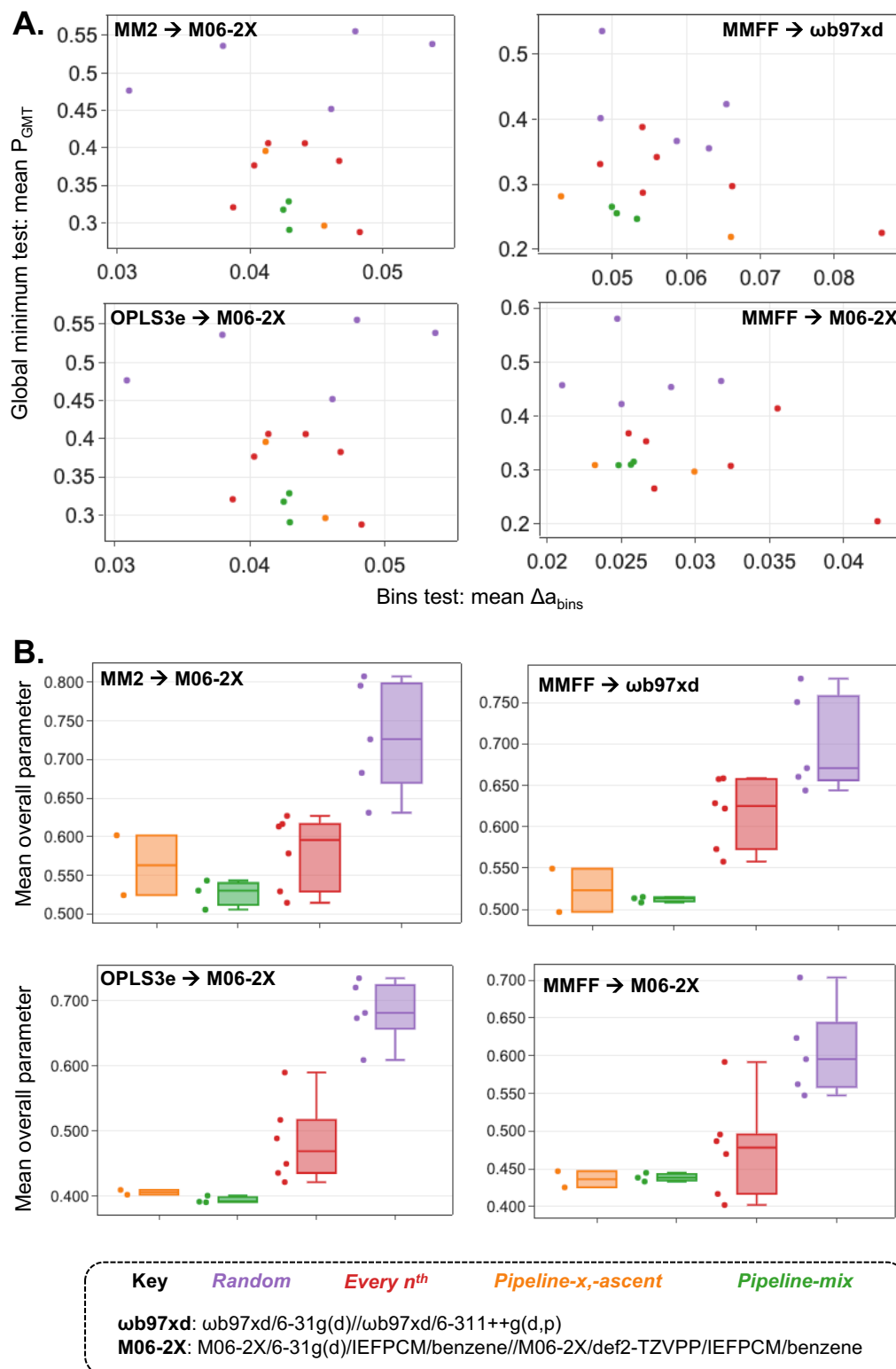


Figure S4.7. Theory levels benchmarking tests: A) The global minimum test (P_{GMT}) and bins test (Δa_{bins}) results are presented in scatter plots. B) The mean overall parameter data is presented as box plots. The data points on the scatter plots and box plots are coloured according to the corresponding priority list generation approach. A reference key is at the bottom of the figure, in which complete descriptions of the theory levels are also given. The tests for the *random approach* have been repeated for 5 times, which contributes to the 5 data points under this category.

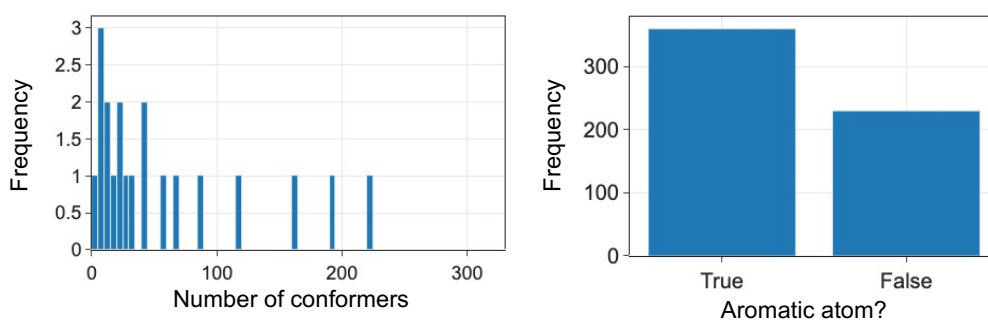


Figure S4.8. An overview of the 20 Grayson molecules. A histogram of the conformer number is given. The bar chart shows that approximately 2/3 of the non-H atoms in the Grayson molecules are identified as aromatic atoms via RDkit.

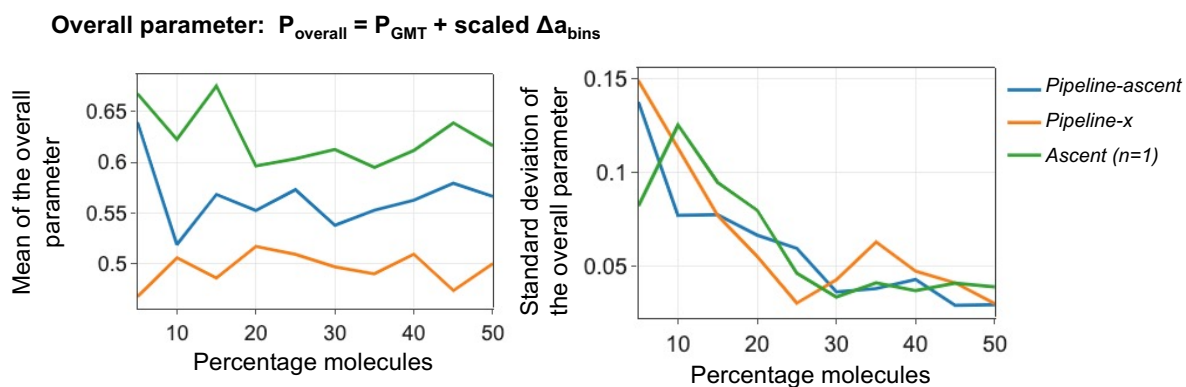


Figure S4.9. Percentage of molecules required from the DFT dataset to reach convergence in the performance parameter. The mean and standard deviation of the overall parameter value vs X% plots are presented.

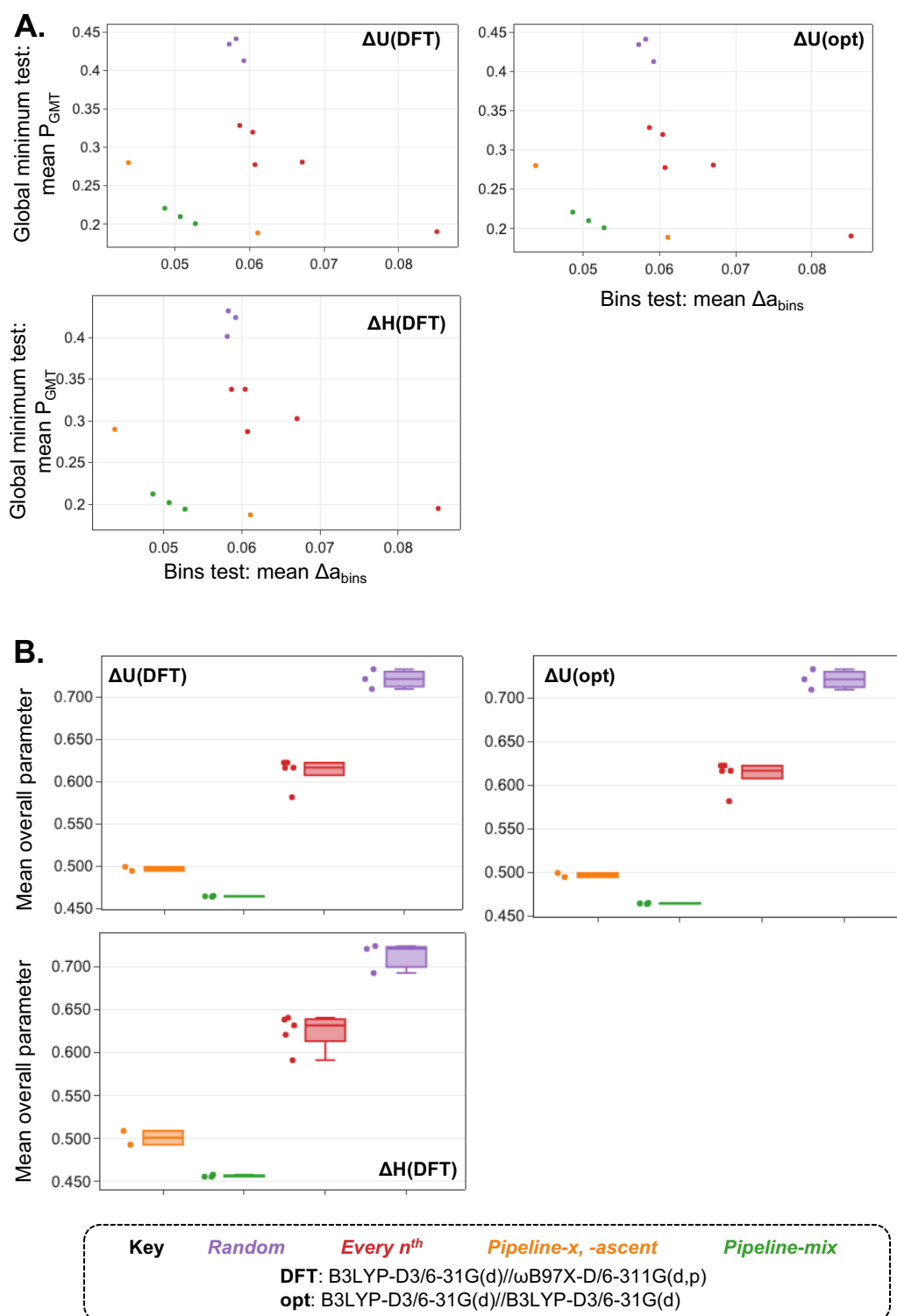


Figure S4.10. Results with different types of energy values. A) The global minimum test (P_{GMT}) and bins test (Δa_{bins}) results are presented in scatter plots. B) The mean overall parameter data is presented as box plots. The data points on the scatter plots and box plots are coloured according to the corresponding priority list generation approach. A reference key is at the bottom of the figure, in which complete descriptions of the theory levels and energy type are also given.

4.F Predict the completion of the re-optimisation process

1. Performance of the local models

The same training procedure was repeated to build a local model with 15 342 data points. The data points come from χ_{new} vs r_{opt} plots from just one set of priority lists, specifically priority lists with the default CONFPASS setting (the *pipeline-mix* approach, $x = 0.8$ and $Q = 0.2$). The local models did not outperform the global models based on results from the cross-validation tests.

The local support vector classification (SVC) model takes more than 4 hours to train with 12 CPU cores computer. Thus, this model was not chosen for building the global model.

Table S4.2: Results of the cross-validation tests on ML models for predicting the completion of the DFT re-optimisation process with the aim of finding the global minimum structure.

	Accuracy	% False Continue
RF	88.07%	8.68%
KNN	87.69%	8.32%
GaussianNB	87.75%	4.91%
LR	89.09%	7.63%

2. Hyperparameter tuning tests

Changes in the hyperparameters did not lead to variations in the performance for the global LR model.

Table S4.3: Hyperparameter tuning cross-validation tests on the global LR model

Max_iter	Random_state	Solver	Penalty	Accuracy	% False Continue
100	0	lbfgs	l2	89.10%	7.60%
200	0	lbfgs	l2	91.44%	2.72%
500	0	lbfgs	l2	88.27%	6.67%
2000	0	lbfgs	l2	88.27%	6.67%
1000	0	lbfgs	l2	89.04%	7.68%
1000	50	lbfgs	l2	89.04%	7.68%
1000	100	lbfgs	l2	89.04%	7.68%
1000	0	newton-cg	l2	89.02%	7.67%
1000	50	newton-cg	l2	89.02%	7.67%
1000	100	newton-cg	l2	89.02%	7.67%
1000	0	saga	l2	89.11%	7.63%
1000	50	saga	l2	89.11%	7.63%
1000	100	saga	l2	89.11%	7.63%

3. Evaluation of the global LR model

Table S4.4. The performance of the global LR model on descriptors generated based on other priority list generation approaches. The rigorous cross-validation methodology mentioned in the main text was followed. (ie For each molecule, relevant data points associated with the molecule, which corresponds to less than 1% of the entire dataset, was taken out. Models were trained on the filtered dataset.)

Input descriptor array			
Priority list generation method	parameter	Accuracy	% False Continue
<i>every n^{th}</i>	$n = 1$	89.29%	5.61%
<i>every n^{th}</i>	$n = 2$	74.33%	22.73%
<i>every n^{th}</i>	$n = 3$	74.23%	23.17%
<i>every n^{th}</i>	$n = 4$	71.00%	27.19%
<i>pipeline-x</i>	$x = 0.8$	89.61%	5.19%
<i>pipeline-ascent</i>	-	83.35%	12.68%
<i>pipeline-mix</i>	$x = 0.8; Q = 0.15$	87.43%	7.79%
<i>pipeline-mix</i>	$x = 0.8; Q = 0.20$ (default for CONFPASS)	88.27%	6.67%
<i>pipeline-mix</i>	$x = 0.8; Q = 0.25$	88.23%	6.62%

4. Boltzmann-weighted property calculations

The ML models were further evaluated from the perspective of Boltzmann distribution.

$r_{\text{opt,conf}=c}$ is the r_{opt} at which %Conf no longer deviates below c in a %Conf vs r_{opt} plot (see Figure S4.11 for an example). The sum of the Boltzmann distribution population percentage for all the re-optimised conformers with a unique structure (p_{total}) was calculated at $r_{\text{opt,conf}=c}$.

For $c = 60\%$ to $c = 90\%$, $r_{\text{opt,conf}=c}$ was calculated and the corresponding p_{total} for 150 molecules using the data from the evaluation process. The mean values were calculated, and results are presented in Table S4.5.

Table S4.5. Boltzmann-weighted property calculation results: $r_{\text{opt,conf}=c}$ and the corresponding p_{total} were found for the 150 molecules. The data are from the evaluation of the global LR model. The table present the mean values of the $r_{\text{opt,conf}=c}$ and p_{total} for $c = 60\%$ to $c = 90\%$. The setting for temperature is 298.15 K.

c (%)	Mean $r_{\text{opt,conf}=c}$	Mean p_{total}
60	0.18	0.67
65	0.20	0.69
70	0.23	0.72
75	0.26	0.74
80	0.30	0.77
85	0.36	0.80
90	0.47	0.86

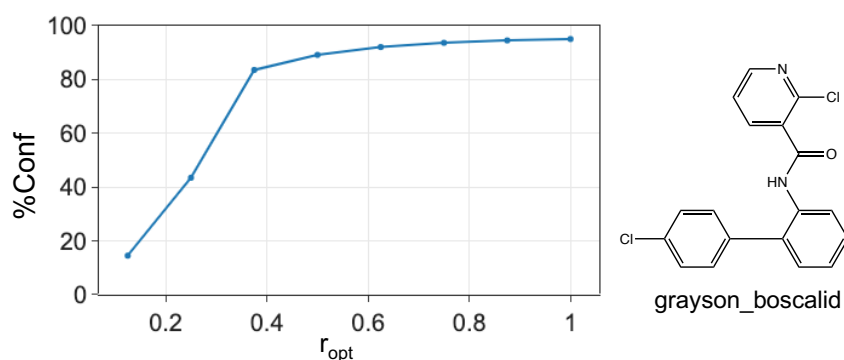


Figure S4.11. %Conf vs r_{opt} plot for grayson_boscalid

4.G Comparison with CREST

1. Conformational searching with CREST^{46,47}

Conformational searching calculations were performed on a selection of 70 molecules (the CREST-CS dataset, Figure S4.14) with CREST using the default setting (method: GFN2-xTB; conformational searching algorithm: iMTD-GC; energy cut-off: 6 kcal mol⁻¹, RMSD cut-off: 0.125 Å).

On average, more conformer structures were found with CREST compared to MacroModel (Figure S4.12). The difference in the results may be due to the difference in the MMFF and the GFN2-xTB potential energy surface (PES). For example, semi-empirical methods are better at capturing potential hydrogen bonding interactions than force field methods. Many unique CREST conformer structures feature H bonds that have not been identified with MacroModel as these structures do not correspond to a minimum on the MMFF PES. Secondly, the structural similarity check in the CREST program might not have taken rotations of symmetrical groups, such as methyl groups, into account. For example, conformer 41 and 42 of 'omegacsd_ABHYTZ_het' are visually identical but differ by the rotation of the methyl group (Figure S4.15).

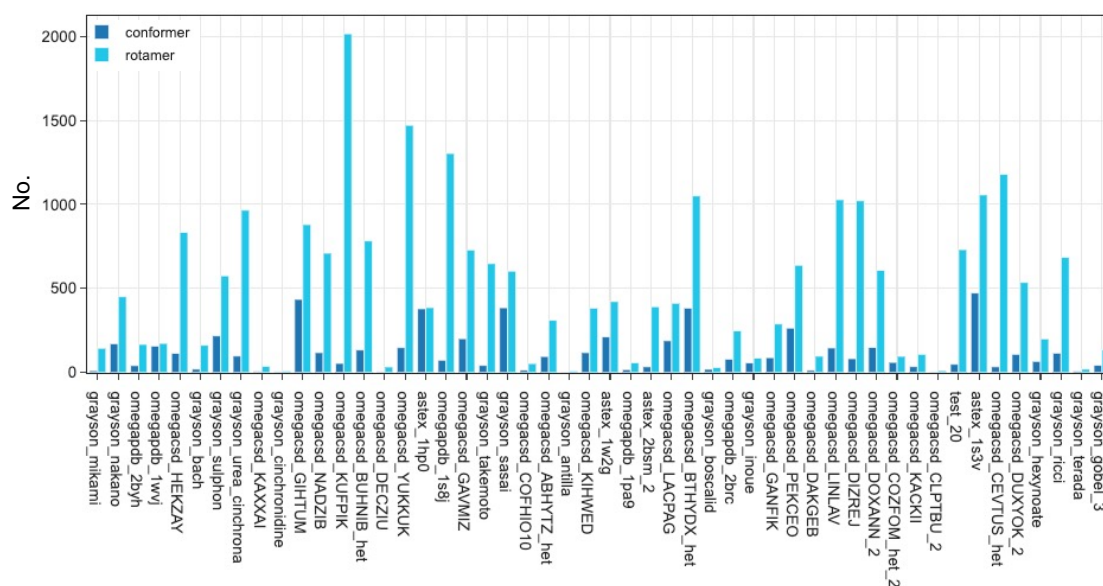


Figure S4.13. The frequency histogram compares the number of conformers and the number of rotamers identified by CREST conformational searches for 50 randomly selected molecules from the CREST-CS dataset.

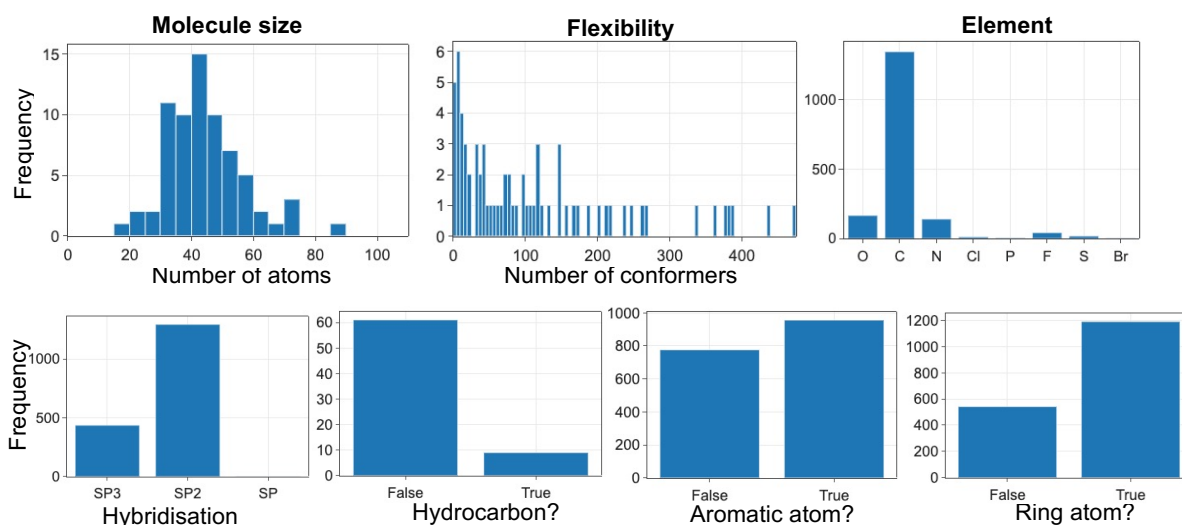


Figure S4.14. The profile of the 70 CREST-CS dataset molecules.

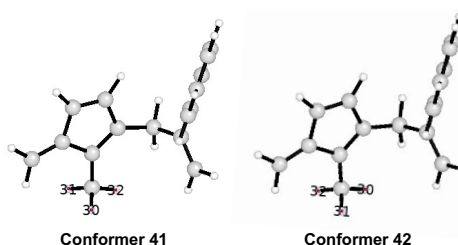


Figure S4.15. Conformer 41 and 42 of 'omegascsd_ABHYTZ_het': The index of the H atoms in the methyl group is labelled.

2. Re-optimisation at the DFT level

20 molecules were randomly selected from the CREST-CS dataset. To evaluate the effectiveness of the CONFPASS pipeline and LR model with the evaluation methods, the procedure in Figure 4.12 of the main text was followed with the CREST conformational searching results of the 20 molecules as the starting point. All GFN2-xTB conformer structures from the CREST conformational searches were re-optimised at the DFT level (*ie* the CREST-DFT dataset): ω b97xd/6-311g(d,p)//B3LYP-D3/6-31g(d). The conformer clusters at the DFT level and the DFT lists were derived.

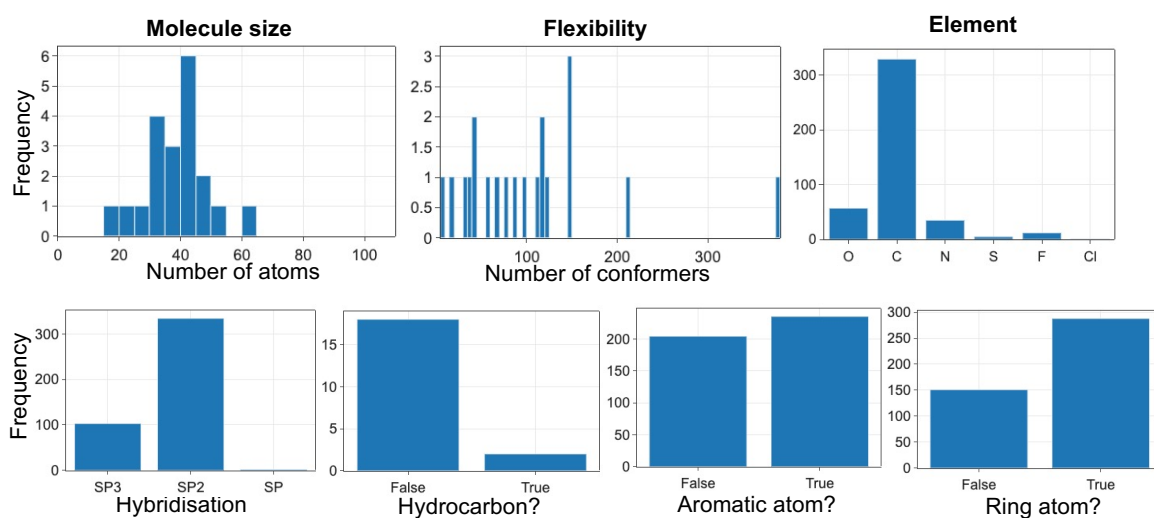


Figure S4.16. The profile of the 20 chosen CREST dataset molecules where DFT re-optimisations were performed on all identified conformers.

3. Performance

All the evaluations below are conducted using Gibbs free energy values with the CREST-DFT dataset (GFN2-xTB $\rightarrow$ ω b97xd/6-311g(d,p)//B3LYP-D3/6-31g(d) optimised conformer structures)

Duplicate conformers

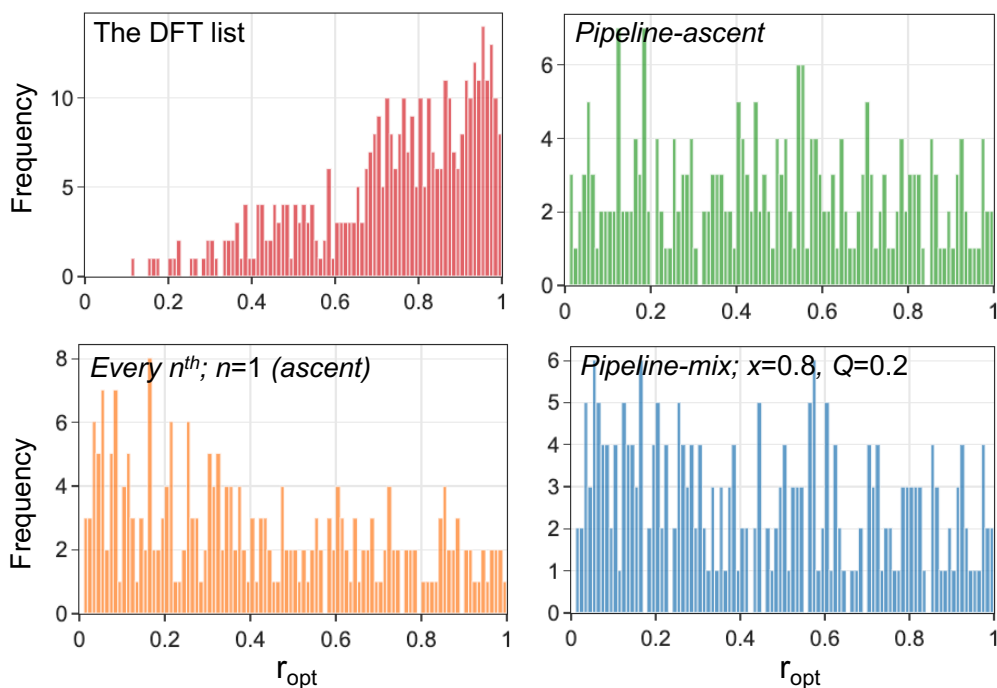


Figure S4.17. The histograms of duplicate conformers appearance frequency over r_{opt} for the *pipeline-ascent* approach, *pipeline-mix* approach, *ascent* approach and the ideal scenario with the DFT list. The setting for the *pipeline-mix* approach is $x = 0.8$ and $Q = 0.2$.

Evaluation on the ML model for predicting the completion of the re-optimisation process

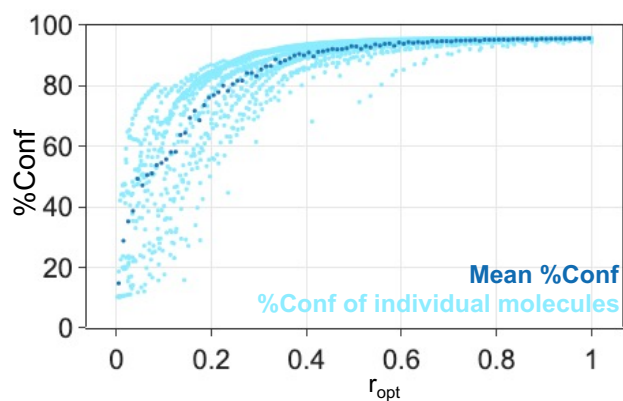


Figure S4.18. The percentage confidence (%Conf) vs r_{opt} plots for the 20 molecules in the CREST DFT dataset. The predictions and %Conf values come from the global LR model in the CONFPASS package. These data points were separated into 100 bins of equal size according to the r_{opt} values. The mean %Conf was found for each bin. The result has been included in the above plot as dark blue points. The priority list generating method is indicated on the plot.

Key terms and their definitions for Table S4.6-S4.8

$r_{\text{opt,conf}=c}$: the r_{opt} at which %Conf no longer deviates below c in a %Conf vs r_{opt} plot

ρ_{total} : the sum of the Boltzmann distribution population percentage for all the re-optimised conformers with a unique structure at 298.15 K

ΔG_{gl} : the ΔG of the lowest energy re-optimised conformer from the group compared to all DFT structures of the molecule (*ie* all re-optimised structures from MacroModel and CREST).

Table S4.6. Performance comparisons between CREST and MacroModel (**performance without CONFPASS, *ie* all conformers re-optimised at the DFT level**): DFT structures re-optimised from MacroModel and Crest outputs were considered together when making the below calculations. ρ_{total} values were calculated at $r_{\text{opt}}=1$ considering all re-optimised structures from MacroModel and CREST conformational searches. A low ρ_{total} implies that key low energy structures found with the other conformational searching method have been missed out with the corresponding searching method. In most cases, CREST and MacroModel lead to the same global minimum structure at the DFT level after re-optimisation (*ie* $\Delta G_{\text{gl}} = 0 \text{ kcal mol}^{-1}$). The scaled CUP time values required for re-optimising all conformers from the conformational searching output are given below.

Molecule Name	ρ_{total} at $r_{\text{opt}}=1$		ΔG_{gl} (kcal mol ⁻¹)		Scaled CPU time	
	MacroModel	CREST	MacroModel	CREST	MacroModel	CREST
omegapdb_1h2k	1.00	1.00	0.0	0.0	1847	314
grayson_inoue	0.99	1.00	0.0	0.0	13605	14135
grayson_ricci	0.99	0.78	0.0	0.0	77169	59567
omegacsd_GANFIK	0.83	0.91	0.0	0.0	17924	23812
omegacsd_LINLAV	0.90	0.96	0.0	0.0	25038	97112
omegapdb_2brc	1.00	0.23	0.0	0.0	39118	58287
astex_1w2g	0.97	0.97	0.0	0.0	6200	30578
omegapdb_2byh	0.96	0.51	0.0	0.0	16000	11648
omegacsd_YUKKUK	0.66	1.00	0.0	0.0	23217	42673
grayson_hexynoate	0.96	0.43	0.0	0.1	8261	7827
grayson_gobel	0.31	1.00	0.6	0.0	8822	34530
Total					237202	380484
Ratio					1	1.60

Table S4.7. Performance comparisons between CREST and MacroModel (**performance with CONFPASS**, ie the incomplete set of re-optimised structures when %Conf no longer deviates below 90% based on the LR model following the priority list from the default setting of CONFPASS). The ρ_{total} and ΔG_{gl} at the corresponding r_{opt} value are reported below. The scaled CPU time below refers to the time required for re-optimising the incomplete set of re-optimised structures when %Conf no longer deviates below 90% following the priority list from CONFPASS with the default setting.

Molecule Name	r_{opt} when %Conf no longer deviates below 90% ($r_{\text{opt,conf}=90\%}$)		ρ_{total}		ΔG_{gl} (kcal mol ⁻¹)	
	MacroM odel	CREST	Macro Model	CREST	MacroM odel	CREST
omegapdb_1h2k	0.39	0.43	1.00	0.87	0.0	0.0
grayson_inoue	0.53	0.46	1.00	1.00	0.0	0.0
grayson_ricci	0.31	0.34	0.89	0.93	0.0	0.0
omegacsd_GANFIK	0.37	0.41	0.99	0.93	0.0	0.0
omegacsd_LINLAV	0.40	0.41	0.98	0.92	0.0	0.0
omegapdb_2brc	0.45	0.52	0.60	0.76	0.0	0.0
astex_1w2g	0.45	0.29	0.79	0.87	0.0	0.0
omegapdb_2byh	0.50	0.54	0.99	1.00	0.0	0.0
omegacsd_YUKKUK	0.61	0.34	0.95	0.89	0.0	0.0
grayson_hexynoate	0.50	0.41	0.41	0.98	0.6	0.0
grayson_gobel	0.54	0.67	0.72	1.00	0.0	0.0

Molecule Name	Total number of conformers from conformational searches		The corresponding number of conformers at $r_{\text{opt,conf}=90\%}$		Scaled CPU time	
	MacroM odel	CREST	Macro Model	CREST	MacroM odel	CREST
omegapdb_1h2k	51	7	20	3	666	150
grayson_inoue	59	57	31	26	7162	6155
grayson_ricci	223	115	70	39	25425	20211
omegacsd_GANFIK	82	88	30	36	6624	9552
omegacsd_LINLAV	40	146	16	60	10700	43999
omegapdb_2brc	155	79	69	41	16648	29117
astex_1w2g	60	212	27	61	2921	7634
omegapdb_2byh	72	41	36	22	8730	6593
omegacsd_YUKKUK	54	148	33	50	13001	14442
grayson_hexynoate	66	66	33	27	3121	3424
grayson_gobel	13	43	7	29	5533	24982
Total					100529	166260
Ratio					1	1.65

Table S4.8. Boltzmann-weighted property calculation results for the 20 molecules in the CREST-DFT dataset: The data are from the evaluation of the global LR model. The table present the mean values of the $r_{\text{opt,conf=c}}$ and ρ_{total} for $c = 60\%$ to $c = 90\%$. The setting for temperature is 298.15 K.

c (%)	Mean $r_{\text{opt,conf=c}}$	Mean ρ_{total}
60	0.14	0.61
65	0.17	0.67
70	0.21	0.72
75	0.24	0.76
80	0.28	0.79
85	0.33	0.84
90	0.43	0.90

Appendix to Chapter 5

5.A Data availability

1. Structural data

The structural information is available in the Cambridge Apollo Repository:

<https://doi.org/10.17863/CAM.96901>

The key structures are included in the 'SI_key_structure_cobi.zip' file as opt+freq or opt Gaussian calculation output files. The optimisations were conducted at the B3LYP-D3/6-31G(d) level of theory.

2. VRAI-selectivity extension: VRAI-multi

The input to VRAI-multi is the path to the folder that contains the required files for the VRAI-selectivity analyses. For the script to correctly identify and process the files, the following file naming formats must be adopted:

- Geometry of the first TS in mol file must have a 'TS1.mol' suffix.
- Gaussian16 frequency calculation output file for the first TS must adopt the same filename as the corresponding mol file but with a '.out' suffix.
 - o Example: *ol17ryu_e_int2_SSR_11_TS1.mol* and *ol17ryu_e_int2_SSR_11_TS1.out*
- Geometry of the intermediate in .mol file must have a 'int.mol' suffix.
- Gaussian16 frequency calculation output file for the intermediate must adopt the same filename as the corresponding mol file but with a '.out' suffix.
 - o Example: *ol17ryu_e_int2_SSR_6_int.mol* and *ol17ryu_e_int2_SSR_6_int.out*
- Gaussian16 frequency calculation output files for the second TS must have a 'TS2.out' suffix. (*optional*)
- Geometry of the product in .mol file must adopt the same filename as the corresponding TS2 with an additional 'prod.mol' as the suffix.
 - o Example: there are three different possible products via pathways that share a common intermediate
ol17ryu_e_int2_SSR_C_1_TS2_prod.mol
ol17ryu_e_int2_SSR_C_1_TS2.out
ol17ryu_e_int2_SSR_H_2_TS2_prod.mol
ol17ryu_e_int2_SSR_H_2_TS2.out
ol17ryu_e_int2_SSR_O_1_TS2_prod.mol
ol17ryu_e_int2_SSR_O_1_TS2.out
- Single point energy calculation files must adopt the same filename as the corresponding TS1, int and TS2 structures but with a '_spe.out' suffix. (*optional*)
 - o Example:
ol17ryu_e_int2_SSR_O_1_TS2_spe.out for *ol17ryu_e_int2_SSR_O_1_TS2.out*

The designated folder must contain a minimum of one set of TS1 files, one set of INT files and two product mol files. It is acceptable to include additional sets of products or TS1 files, but the number of intermediate files in the folder should not exceed one set (*ie* a mol and out file).

5.B Calculation breakdowns

Notes on the result presentations: Unless otherwise specified, the following notations apply to all the tables below.

- Result tables for 'TST only' predictions: The selectivity is controlled kinetically by TS1 and TS2. $\Delta\Delta G^\ddagger$ is in kcal mol⁻¹ at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) level of theory. %(TS1) and %(TS2) corresponds to the percentage population of the TS1 and TS2. %(Pathway) values were calculated as a product of %(TS1) and %(TS2).
- Result tables for 'with VRAI' predictions: %(TS1) refers to the percentage population of the TS1. %(VRAI) is the predicted product percentage from VRAI-selectivity calculations. The frequency analyses required for VRAI-selectivity calculations were conducted at the B3LYP-D3/6-31G(d) level of theory. In the VRAI-selectivity calculation, the given TS1 corresponds to the first transition state and the product structures come from the QRC calculations of the given TS2. %(Pathway) values were calculated as a product of %(TS1) and %(VRAI).

RMSD calculations

Root-mean-square deviation (RMSD) calculations were performed with the GetBestRMS function under the rdkit.Chem.rdMolAlign module on optimised TSs to identify the duplicate structures, which should also be also degenerate in energy. The pair of conformers were considered as having the same structure if the RMSD value < 0.005 Å. The duplicate structures were removed from the dataset before calculating product percentages.

The ketone-selective reaction with Cat-B

TST only:

Table S5.1. Calculation breakdowns for the predicted product percentages of the ketone-selective reaction with Cat-B based on TST only. Adding 'ol17ryu_b_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the structure. ($\Delta\Delta G^\ddagger$ Unit: kcal mol⁻¹)

TS1	%(TS1)	$\Delta\Delta G^\ddagger$ (TS1)	TS2	%(TS2)	$\Delta\Delta G^\ddagger$ (TS2)	%(Pathway)	Product
RR_2_TS1	13.2%	0.72	RR_C_1_TSe	4.9%	1.11	0.7%	R aldehyde
RR_2_TS1	13.2%	0.72	RR_C_2_TSe	3.2%	1.27	0.4%	R aldehyde
RR_2_TS1	13.2%	0.72	RR_C_3_TSe	2.4%	1.39	0.3%	R aldehyde
RR_2_TS1	13.2%	0.72	RR_H_1_TSe	85.8%	0.00	11.3%	S ketone
RR_2_TS1	13.2%	0.72	RR_H_2_TSe	3.7%	1.22	0.5%	S ketone
RR_8_TS1	2.0%	1.44	RR_C_1_TSe	4.9%	1.11	0.1%	R aldehyde
RR_8_TS1	2.0%	1.44	RR_C_2_TSe	3.2%	1.27	0.1%	R aldehyde
RR_8_TS1	2.0%	1.44	RR_C_3_TSe	2.4%	1.39	0.0%	R aldehyde
RR_8_TS1	2.0%	1.44	RR_H_1_TSe	85.8%	0.00	1.7%	S ketone
RR_8_TS1	2.0%	1.44	RR_H_2_TSe	3.7%	1.22	0.1%	S ketone
RS_1_TS1_2	83.4%	0.00	RS_C_1_TSe	43.5%	0.00	36.3%	S aldehyde
RS_1_TS1_2	83.4%	0.00	RS_C_2_TSe	16.0%	0.39	13.3%	S aldehyde
RS_1_TS1_2	83.4%	0.00	RS_C_3_TSe	2.3%	1.15	1.9%	S aldehyde
RS_1_TS1_2	83.4%	0.00	RS_H_1_TSe	6.9%	0.71	5.8%	R ketone
RS_1_TS1_2	83.4%	0.00	RS_H_2_TSe	5.6%	0.80	4.7%	R ketone
RS_1_TS1_2	83.4%	0.00	RS_H_3_TSe	1.4%	1.32	1.2%	R ketone
RS_1_TS1_2	83.4%	0.00	RS_H_4_TSe	9.5%	0.59	8.0%	R ketone
RS_1_TS1_2	83.4%	0.00	RS_O_1_TSe	5.4%	0.81	4.5%	RR epoxide
RS_1_TS1_2	83.4%	0.00	RS_O_2_TSe	6.6%	0.73	5.5%	RR epoxide
RS_1_TS1_2	83.4%	0.00	RS_O_5_TSe	2.8%	1.07	2.3%	RR epoxide
RS_9_TS1	1.3%	1.60	RS_C_1_TSe	43.5%	0.00	0.6%	S aldehyde
RS_9_TS1	1.3%	1.60	RS_C_2_TSe	16.0%	0.39	0.2%	S aldehyde
RS_9_TS1	1.3%	1.60	RS_C_3_TSe	2.3%	1.15	0.0%	S aldehyde
RS_9_TS1	1.3%	1.60	RS_H_1_TSe	6.9%	0.71	0.1%	R ketone
RS_9_TS1	1.3%	1.60	RS_H_2_TSe	5.6%	0.80	0.1%	R ketone
RS_9_TS1	1.3%	1.60	RS_H_3_TSe	1.4%	1.32	0.0%	R ketone
RS_9_TS1	1.3%	1.60	RS_H_4_TSe	9.5%	0.59	0.1%	R ketone
RS_9_TS1	1.3%	1.60	RS_O_1_TSe	5.4%	0.81	0.1%	RR epoxide
RS_9_TS1	1.3%	1.60	RS_O_2_TSe	6.6%	0.73	0.1%	RR epoxide
RS_9_TS1	1.3%	1.60	RS_O_5_TSe	2.8%	1.07	0.0%	RR epoxide

With VRAI:

Table S5.2. Calculation breakdowns for the predicted product percentages of the ketone-selective reaction with Cat-B incorporating VRAI-selectivity calculations (see Figure 5 in the main text for detail). Adding 'ol17ryu_b_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the relevant structure.

TS1	%(TS1)	TS2	%(VRAI)	%(Pathway)	Product
RR_2_TS1	13.2%	RR_H_1_TSe	89.7%	11.8%	S ketone
RR_2_TS1	13.2%	RR_C_1_TSe	10.3%	1.4%	R aldehyde
RR_8_TS1	2.0%	RR_C_1_TSe	96.1%	2.0%	R aldehyde
RR_8_TS1	2.0%	RR_H_1_TSe	3.9%	0.1%	S ketone
RS_1_TS1_2	83.4%	RS_H_4_TSe	96.1%	80.2%	R ketone
RS_1_TS1_2	83.4%	RS_O_2_TSe	3.9%	3.3%	RR epoxide
RS_1_TS1_2	83.4%	RS_C_1_TSe	0.0%	0.0%	S aldehyde
RS_9_TS1	1.3%	RS_O_2_TSe	80.2%	1.1%	RR epoxide
RS_9_TS1	1.3%	RS_H_4_TSe	19.8%	0.3%	R ketone
RS_9_TS1	1.3%	RS_C_1_TSe	0.0%	0.0%	S aldehyde

The ketone-selective reaction with Cat-C

TST only:

Table S5.3. Calculation breakdowns for the predicted product percentages of the ketone-selective reaction with Cat-C based on TST only. Adding 'ol17ryu_c_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the structure. ($\Delta\Delta G^\ddagger$ Unit: kcal mol⁻¹)

TS1	%(TS1)	$\Delta\Delta G^\ddagger$ (TS1)	TS2	%(TS2)	$\Delta\Delta G^\ddagger$ (TS2)	%(Pathway)	Product
SR_9_TS1	53.8%	0.00	SR_C_2_TSe	14.8%	0.52	8.0%	R aldehyde
SR_9_TS2	53.8%	0.00	SR_C_3_TSe	5.3%	0.92	2.8%	R aldehyde
SR_9_TS3	53.8%	0.00	SR_C_4_TSe	0.9%	1.60	0.5%	R aldehyde
SR_9_TS4	53.8%	0.00	SR_H_3_TSe	56.2%	0.00	30.2%	S ketone
SR_9_TS5	53.8%	0.00	SR_H_8_TSe	15.0%	0.51	8.1%	S ketone
SR_9_TS6	53.8%	0.00	SR_H_10_TSe	7.3%	0.79	3.9%	S ketone
SR_9_TS7	53.8%	0.00	SR_H_12_TSe	0.6%	1.79	0.3%	S ketone
SR_e3_TS1	18.4%	0.42	SR_C_2_TSe	14.8%	0.52	2.7%	R aldehyde
SR_e3_TS2	18.4%	0.42	SR_C_3_TSe	5.3%	0.92	1.0%	R aldehyde
SR_e3_TS3	18.4%	0.42	SR_C_4_TSe	0.9%	1.60	0.2%	R aldehyde
SR_e3_TS4	18.4%	0.42	SR_H_3_TSe	56.2%	0.00	10.4%	S ketone
SR_e3_TS5	18.4%	0.42	SR_H_8_TSe	15.0%	0.51	2.8%	S ketone
SR_e3_TS6	18.4%	0.42	SR_H_10_TSe	7.3%	0.79	1.3%	S ketone
SR_e3_TS7	18.4%	0.42	SR_H_12_TSe	0.6%	1.79	0.1%	S ketone
SR_e11_TS1	7.9%	0.74	SR_C_2_TSe	14.8%	0.52	1.2%	R aldehyde
SR_e11_TS2	7.9%	0.74	SR_C_3_TSe	5.3%	0.92	0.4%	R aldehyde
SR_e11_TS3	7.9%	0.74	SR_C_4_TSe	0.9%	1.60	0.1%	R aldehyde
SR_e11_TS4	7.9%	0.74	SR_H_3_TSe	56.2%	0.00	4.4%	S ketone
SR_e11_TS5	7.9%	0.74	SR_H_8_TSe	15.0%	0.51	1.2%	S ketone
SR_e11_TS6	7.9%	0.74	SR_H_10_TSe	7.3%	0.79	0.6%	S ketone
SR_e11_TS7	7.9%	0.74	SR_H_12_TSe	0.6%	1.79	0.0%	S ketone
SR_e13_TS1	7.0%	0.79	SR_C_2_TSe	14.8%	0.52	1.0%	R aldehyde
SR_e13_TS1	7.0%	0.79	SR_C_3_TSe	5.3%	0.92	0.4%	R aldehyde
SR_e13_TS1	7.0%	0.79	SR_C_4_TSe	0.9%	1.60	0.1%	R aldehyde
SR_e13_TS1	7.0%	0.79	SR_H_3_TSe	56.2%	0.00	3.9%	S ketone
SR_e13_TS1	7.0%	0.79	SR_H_8_TSe	15.0%	0.51	1.0%	S ketone
SR_e13_TS1	7.0%	0.79	SR_H_10_TSe	7.3%	0.79	0.5%	S ketone
SR_e13_TS1	7.0%	0.79	SR_H_12_TSe	0.6%	1.79	0.0%	S ketone
SS_13_TS1	12.9%	0.56	SS_C_2_TSe	30.1%	0.33	3.9%	S aldehyde
SS_13_TS1	12.9%	0.56	SS_C_3_TSe	69.9%	0.00	9.0%	S aldehyde

With VRAI:

Table S5.4. Calculation breakdowns for the predicted product percentages of the ketone-selective reaction with Cat-C incorporating VRAI-selectivity calculations. Adding 'ol17ryu_c_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the relevant structure.

TS1	%(TS1)	TS2	%(VRAI)	%(Pathway)	Product
SR_9_TS1	53.8%	SR_H_3_TSe	78.7%	42.4%	S ketone
SR_9_TS1	53.8%	SR_O_8_TSe	21.0%	11.3%	SS epoxide
SR_9_TS1	53.8%	SR_C_2_TSe	0.3%	0.2%	R aldehyde
SR_e3_TS1	18.4%	SR_C_2_TSe	37.6%	6.9%	R aldehyde
SR_e3_TS1	18.4%	SR_H_3_TSe	32.8%	6.0%	S ketone
SR_e3_TS1	18.4%	SR_O_8_TSe	29.7%	5.5%	SS epoxide
SR_e11_TS1	7.9%	SR_H_3_TSe	36.2%	2.9%	S ketone
SR_e11_TS1	7.9%	SR_C_2_TSe	32.9%	2.6%	R aldehyde
SR_e11_TS1	7.9%	SR_O_8_TSe	30.9%	2.4%	SS epoxide
SR_e13_TS1	7.0%	SR_H_3_TSe	34.5%	2.4%	S ketone
SR_e13_TS1	7.0%	SR_C_2_TSe	34.2%	2.4%	R aldehyde
SR_e13_TS1	7.0%	SR_O_8_TSe	31.2%	2.2%	SS epoxide
SS_13_TS1	12.9%	SS_H_1_TSe	74.1%	9.5%	R ketone
SS_13_TS1	12.9%	SS_C_3_TSe	25.9%	3.3%	S aldehyde

The ketone-selective reaction with Cat-D

TST only:

Table S5.5. Calculation breakdowns for the predicted product percentages of the ketone-selective reaction with Cat-D based on TST only. Adding 'ol17ryu_e_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the structure. ($\Delta\Delta G^\ddagger$ Unit: kcal mol⁻¹)

TS1	%(TS1)	$\Delta\Delta G^\ddagger$ (TS1)	TS2	%(TS2)	$\Delta\Delta G^\ddagger$ (TS2)	%(Pathway)	Product
SR_3_TS1	2.2%	1.27	SR_C_1_TSe	3.7%	1.16	0.1%	R aldehyde
SR_3_TS1	2.2%	1.27	SR_C_2_TSe	3.4%	1.20	0.1%	R aldehyde
SR_3_TS1	2.2%	1.27	SR_C_3_TSe	1.0%	1.67	0.0%	R aldehyde
SR_3_TS1	2.2%	1.27	SR_H_1_TSe	2.2%	1.37	0.0%	S ketone
SR_3_TS1	2.2%	1.27	SR_H_2_TSe	75.4%	0.00	1.7%	S ketone
SR_3_TS1	2.2%	1.27	SR_H_3_TSe	0.8%	1.74	0.0%	S ketone
SR_3_TS1	2.2%	1.27	SR_H_4_TSe	10.0%	0.78	0.2%	S ketone
SR_3_TS1	2.2%	1.27	SR_H_6_TSe	2.1%	1.39	0.0%	S ketone
SR_3_TS1	2.2%	1.27	SR_H_7_TSe	1.3%	1.57	0.0%	S ketone
SR_9_TS1	2.3%	1.26	SR_C_1_TSe	3.7%	1.16	0.1%	R aldehyde
SR_9_TS1	2.3%	1.26	SR_C_2_TSe	3.4%	1.20	0.1%	R aldehyde
SR_9_TS1	2.3%	1.26	SR_C_3_TSe	1.0%	1.67	0.0%	R aldehyde
SR_9_TS1	2.3%	1.26	SR_H_1_TSe	2.2%	1.37	0.1%	S ketone
SR_9_TS1	2.3%	1.26	SR_H_2_TSe	75.4%	0.00	1.7%	S ketone
SR_9_TS1	2.3%	1.26	SR_H_3_TSe	0.8%	1.74	0.0%	S ketone
SR_9_TS1	2.3%	1.26	SR_H_4_TSe	10.0%	0.78	0.2%	S ketone
SR_9_TS1	2.3%	1.26	SR_H_6_TSe	2.1%	1.39	0.0%	S ketone
SR_9_TS1	2.3%	1.26	SR_H_7_TSe	1.3%	1.57	0.0%	S ketone
SR_11_TS1	59.6%	0.00	SR_C_1_TSe	3.7%	1.16	2.2%	R aldehyde
SR_11_TS1	59.6%	0.00	SR_C_2_TSe	3.4%	1.20	2.0%	R aldehyde
SR_11_TS1	59.6%	0.00	SR_C_3_TSe	1.0%	1.67	0.6%	R aldehyde
SR_11_TS1	59.6%	0.00	SR_H_1_TSe	2.2%	1.37	1.3%	S ketone
SR_11_TS1	59.6%	0.00	SR_H_2_TSe	75.4%	0.00	44.9%	S ketone
SR_11_TS1	59.6%	0.00	SR_H_3_TSe	0.8%	1.74	0.5%	S ketone
SR_11_TS1	59.6%	0.00	SR_H_4_TSe	10.0%	0.78	6.0%	S ketone
SR_11_TS1	59.6%	0.00	SR_H_6_TSe	2.1%	1.39	1.3%	S ketone
SR_11_TS1	59.6%	0.00	SR_H_7_TSe	1.3%	1.57	0.8%	S ketone
SR_13_TS1	4.9%	0.97	SR_C_1_TSe	3.7%	1.16	0.2%	R aldehyde
SR_13_TS1	4.9%	0.97	SR_C_2_TSe	3.4%	1.20	0.2%	R aldehyde
SR_13_TS1	4.9%	0.97	SR_C_3_TSe	1.0%	1.67	0.1%	R aldehyde
SR_13_TS1	4.9%	0.97	SR_H_1_TSe	2.2%	1.37	0.1%	S ketone
SR_13_TS1	4.9%	0.97	SR_H_2_TSe	75.4%	0.00	3.7%	S ketone
SR_13_TS1	4.9%	0.97	SR_H_3_TSe	0.8%	1.74	0.0%	S ketone
SR_13_TS1	4.9%	0.97	SR_H_4_TSe	10.0%	0.78	0.5%	S ketone
SR_13_TS1	4.9%	0.97	SR_H_6_TSe	2.1%	1.39	0.1%	S ketone
SR_13_TS1	4.9%	0.97	SR_H_7_TSe	1.3%	1.57	0.1%	S ketone
SR_c9_TS1	22.0%	0.39	SR_C_1_TSe	3.7%	1.16	0.8%	R aldehyde
SR_c9_TS1	22.0%	0.39	SR_C_2_TSe	3.4%	1.20	0.7%	R aldehyde
SR_c9_TS1	22.0%	0.39	SR_C_3_TSe	1.0%	1.67	0.2%	R aldehyde

SR_c9_TS1	22.0%	0.39	SR_H_1_TSe	2.2%	1.37	0.5%	S ketone
SR_c9_TS1	22.0%	0.39	SR_H_2_TSe	75.4%	0.00	16.6%	S ketone
SR_c9_TS1	22.0%	0.39	SR_H_3_TSe	0.8%	1.74	0.2%	S ketone
SR_c9_TS1	22.0%	0.39	SR_H_4_TSe	10.0%	0.78	2.2%	S ketone
SR_c9_TS1	22.0%	0.39	SR_H_6_TSe	2.1%	1.39	0.5%	S ketone
SR_c9_TS1	22.0%	0.39	SR_H_7_TSe	1.3%	1.57	0.3%	S ketone
SR_x1_TS1	2.2%	1.27	SR_C_1_TSe	3.7%	1.16	0.1%	R aldehyde
SR_x1_TS1	2.2%	1.27	SR_C_2_TSe	3.4%	1.20	0.1%	R aldehyde
SR_x1_TS1	2.2%	1.27	SR_C_3_TSe	1.0%	1.67	0.0%	R aldehyde
SR_x1_TS1	2.2%	1.27	SR_H_1_TSe	2.2%	1.37	0.0%	S ketone
SR_x1_TS1	2.2%	1.27	SR_H_2_TSe	75.4%	0.00	1.7%	S ketone
SR_x1_TS1	2.2%	1.27	SR_H_3_TSe	0.8%	1.74	0.0%	S ketone
SR_x1_TS1	2.2%	1.27	SR_H_4_TSe	10.0%	0.78	0.2%	S ketone
SR_x1_TS1	2.2%	1.27	SR_H_6_TSe	2.1%	1.39	0.0%	S ketone
SR_x1_TS1	2.2%	1.27	SR_H_7_TSe	1.3%	1.57	0.0%	S ketone
SS_c13_TS1	6.7%	0.85	SS_C_1_TSe	14.4%	0.58	1.0%	S aldehyde
SS_c13_TS1	6.7%	0.85	SS_C_2_TSe	14.4%	0.58	1.0%	S aldehyde
SS_c13_TS1	6.7%	0.85	SS_C_3_TSe	63.8%	0.00	4.3%	S aldehyde
SS_c13_TS1	6.7%	0.85	SS_C_4_TSe	3.5%	1.13	0.2%	S aldehyde
SS_c13_TS1	6.7%	0.85	SS_C_5_TSe	3.9%	1.09	0.3%	S aldehyde

With VRAI:

Table S5.6. Calculation breakdowns for the predicted product percentages of the ketone-selective reaction with Cat-D based on TST and VRAI-selectivity calculations. Adding 'ol17ryu_e_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the relevant structure.

TS1	%(TS1)	TS2	%(VRAI)	%(Pathway)	Product
SR_3_TS1	2.2%	SR_H_2_TSe	91.2%	2.0%	S ketone
SR_3_TS1	2.2%	SR_O_1_TSe	8.8%	0.2%	SS epoxide
SR_3_TS1	2.2%	SR_C_1_TSe	0.0%	0.0%	R aldehyde
SR_9_TS1	2.3%	SR_H_2_TSe	92.5%	2.1%	S ketone
SR_9_TS1	2.3%	SR_O_1_Tse	7.5%	0.2%	SS epoxide
SR_9_TS1	2.3%	SR_C_1_Tse	0.0%	0.0%	R aldehyde
SR_11_TS1	59.6%	SR_H_2_Tse	88.6%	52.8%	S ketone
SR_11_TS1	59.6%	SR_O_1_Tse	10.2%	6.1%	SS epoxide
SR_11_TS1	59.6%	SR_C_1_Tse	1.3%	0.7%	R aldehyde
SR_13_TS1	4.9%	SR_H_2_Tse	94.2%	4.6%	S ketone
SR_13_TS1	4.9%	SR_O_1_Tse	5.8%	0.3%	SS epoxide
SR_13_TS1	4.9%	SR_C_1_Tse	0.0%	0.0%	R aldehyde
SR_c9_TS1	22.0%	SR_H_2_Tse	43.1%	9.5%	S ketone
SR_c9_TS1	22.0%	SR_C_1_Tse	41.7%	9.2%	R aldehyde
SR_c9_TS1	22.0%	SR_O_1_Tse	15.2%	3.3%	SS epoxide
SR_x1_TS1	2.2%	SR_H_2_Tse	91.2%	2.0%	S ketone
SR_x1_TS1	2.2%	SR_O_1_Tse	8.8%	0.2%	SS epoxide
SR_x1_TS1	2.2%	SR_C_1_Tse	0.0%	0.0%	R aldehyde
SS_c13_TS1	6.7%	SS_H_3_Tse	41.5%	2.8%	R ketone
SS_c13_TS1	6.7%	SS_O_1_Tse	37.6%	2.5%	SR epoxide
SS_c13_TS1	6.7%	SS_C_3_Tse	20.9%	1.4%	S aldehyde

The epoxide-selective reaction

TST:

Table S5.7. Calculation breakdowns for the predicted product percentages of the epoxide-selective reaction based on TST only. %(TS3) corresponds to the percentage population of the TS3. %(Pathway) values are calculated as a product of %(TS1), %(TS2) and %(TS3). Adding 'agw21ryu_d_int2_S' to labels given in TS1, TS2 and TS3 column gives the Gaussian 16 output filename of the structure. NA implies that the process is barrierless and hence no corresponding TS structure is obtained via this pathway. ($\Delta\Delta G^\ddagger$ Unit: kcal mol⁻¹)

TS1	%(TS1)	$\Delta\Delta G^\ddagger(\text{TS1})$	TS2	%(TS2)	$\Delta\Delta G^\ddagger(\text{TS2})$	TS3	%(TS3)	$\Delta\Delta G^\ddagger(\text{TS3})$	%(Pathway)	Product
SR_2_TS1	0.8%	1.82	NA	100.0%	NA	NA	100.0%	NA	0.83%	SR epoxide
SR_3_TS1	1.1%	1.71	NA	100.0%	NA	NA	100.0%	NA	1.11%	R aldehyde
SS_10_TS1	2.7%	1.36	SS_Ha_2_TSe	34.8%	0.17	SS_Ha_2_C_TS2_5	99.7%	0.00	0.94%	R aldehyde
SS_10_TS1	2.7%	1.36	SS_Ca_1_TSe	0.1%	2.53	SS_Ca_1_TS2_3	17.9%	0.90	0.00%	S aldehyde
SS_10_TS1	2.7%	1.36	SS_Ca_1_TSe	0.1%	2.53	SS_Ca_1_H_TS2	82.1%	0.00	0.00%	R ketone
SS_10_TS1	2.7%	1.36	SS_Ha_2_TSe	34.8%	0.17	SS_Ha_2_TS2_2	0.3%	3.49	0.00%	S ketone
SS_10_TS1	2.7%	1.36	SS_Oa_2_TSe	11.4%	0.60	NA	100.0%	NA	0.31%	SR epoxide
SS_10_TS1	2.7%	1.36	SS_Ha_3_TSe	53.7%	0.00	NA	100.0%	NA	1.45%	SS epoxide
SS_15_TS1	1.0%	1.74	NA	100.0%	NA	NA	100.0%	NA	1.02%	S aldehyde
SS_50_TS1	3.6%	1.26	SS_Ha_2_TSe	34.8%	0.17	SS_Ha_2_C_TS2_5	99.7%	0.00	1.23%	R aldehyde
SS_50_TS1	3.6%	1.26	SS_Ca_1_TSe	0.1%	2.53	SS_Ca_1_TS2_3	17.9%	0.90	0.00%	S aldehyde
SS_50_TS1	3.6%	1.26	SS_Ca_1_TSe	0.1%	2.53	SS_Ca_1_H_TS2	82.1%	0.00	0.00%	R ketone
SS_50_TS1	3.6%	1.26	SS_Ha_2_TSe	34.8%	0.17	SS_Ha_2_TS2_2	0.3%	3.49	0.00%	S ketone
SS_50_TS1	3.6%	1.26	SS_Oa_2_TSe	11.4%	0.60	NA	100.0%	NA	0.41%	SR epoxide
SS_50_TS1	3.6%	1.26	SS_Ha_3_TSe	53.7%	0.00	NA	100.0%	NA	1.91%	SS epoxide
SS_65_TS1	90.8%	0.00	SS_Ha_2_TSe	34.8%	0.17	SS_Ha_2_C_TS2_5	99.7%	0.00	31.49%	R aldehyde
SS_65_TS1	90.8%	0.00	SS_Ca_1_TSe	0.1%	2.53	SS_Ca_1_TS2_3	17.9%	0.90	0.01%	S aldehyde
SS_65_TS1	90.8%	0.00	SS_Ca_1_TSe	0.1%	2.53	SS_Ca_1_H_TS2	82.1%	0.00	0.06%	R ketone
SS_65_TS1	90.8%	0.00	SS_Ha_2_TSe	34.8%	0.17	SS_Ha_2_TS2_2	0.3%	3.49	0.09%	S ketone
SS_65_TS1	90.8%	0.00	SS_Oa_2_TSe	11.4%	0.60	NA	100.0%	NA	10.35%	SR epoxide
SS_65_TS1	90.8%	0.00	SS_Ha_3_TSe	53.7%	0.00	NA	100.0%	NA	48.79%	SS epoxide

With VRAI:

Table S5.8. Calculation breakdowns for the predicted product percentages of the epoxide-selective reaction based on TST and VRAI-selectivity calculation. Adding 'agw21ryu_d_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the relevant structure. NA implies that the process is barrierless and hence no corresponding TS structure is obtained via this pathway.

TS1	%(TS1)	TS2/TS3	%(VRAI)	%(Pathway)	Product
SR_2_TS1	0.8%	NA	1.0%	0.0%	SR epoxide
SR_3_TS1	1.1%	NA	1.0%	0.0%	R aldehyde
SS_15_TS1	1.0%	NA	1.0%	0.0%	S aldehyde
SS_10_TS1	2.7%	SS_Ha_2_C_TS2_5	34.3%	0.9%	R aldehyde
SS_10_TS1	2.7%	SS_Ca_1_TS2_3	11.6%	0.3%	S aldehyde
SS_10_TS1	2.7%	SS_Ca_1_H_TS2	8.6%	0.2%	R ketone
SS_10_TS1	2.7%	SS_Ha_2_TS2_2	5.2%	0.1%	S ketone
SS_10_TS1	2.7%	SS_Oa_2_TSe	34.3%	0.9%	SR epoxide
SS_10_TS1	2.7%	SS_Ha_3_TSe	6.0%	0.2%	SS epoxide
SS_50_TS1	3.6%	SS_Ha_2_C_TS2_5	35.6%	1.3%	R aldehyde
SS_50_TS1	3.6%	SS_Ca_1_TS2_3	34.3%	1.2%	S aldehyde
SS_50_TS1	3.6%	SS_Ca_1_H_TS2	2.7%	0.1%	R ketone
SS_50_TS1	3.6%	SS_Ha_2_TS2_2	9.3%	0.3%	S ketone
SS_50_TS1	3.6%	SS_Oa_2_TSe	2.7%	0.1%	SR epoxide
SS_50_TS1	3.6%	SS_Ha_3_TSe	15.4%	0.5%	SS epoxide
SS_65_TS1	90.8%	SS_Ha_2_C_TS2_5	36.0%	32.7%	R aldehyde
SS_65_TS1	90.8%	SS_Ca_1_TS2_3	8.3%	7.6%	S aldehyde
SS_65_TS1	90.8%	SS_Ca_1_H_TS2	7.9%	7.2%	R ketone
SS_65_TS1	90.8%	SS_Ha_2_TS2_2	2.6%	2.3%	S ketone
SS_65_TS1	90.8%	SS_Oa_2_TSe	41.7%	37.8%	SR epoxide
SS_65_TS1	90.8%	SS_Ha_3_TSe	3.5%	3.2%	SS epoxide

The aldehyde-selective reaction

TST only:

Table S5.9. Calculation breakdowns for the predicted product percentages of the aldehyde-selective reaction based on TST only. Adding 'jacs13ryu2_d_int2_S' to labels given in TS1 and TS2 column gives the Gaussian 16 output filename of the structure. Some files may have additional suffixes. ($\Delta\Delta G^\ddagger$ Unit: kcal mol⁻¹)

Entry	TS1	%(TS1)	$\Delta\Delta G^\ddagger$ (TS1)	TS2	%(TS2)	$\Delta\Delta G^\ddagger$ (TS2)	%(Pathway)	Product
0	RR_1_TS1_2	89.7%	0.00	RR_H_1_TSe	95.4%	0.00	85.5%	R ketone
1	RR_1_TS1_2	89.7%	0.00	RR_H_3_TSe	4.6%	1.17	4.1%	R ketone
2	RR_3_TS1	1.4%	1.60	RR_H_1_TSe	95.4%	0.00	1.4%	R ketone
3	RR_3_TS1	1.4%	1.60	RR_H_3_TSe	4.6%	1.17	0.1%	R ketone
4	RS_2_TS1	2.3%	1.43	RS_C_1_TSe	73.3%	0.00	1.7%	R aldehyde
5	RS_2_TS1	2.3%	1.43	RS_C_2_TSe	13.0%	0.67	0.3%	R aldehyde
6	RS_2_TS1	2.3%	1.43	RS_C_3_TSe	9.3%	0.80	0.2%	R aldehyde
7	RS_2_TS1	2.3%	1.43	RS_C_4_TSe	1.9%	1.42	0.0%	R aldehyde
8	RS_2_TS1	2.3%	1.43	RS_H_1_TSe	2.6%	1.29	0.1%	S ketone
9	RS_3_TS1	1.6%	1.57	RS_C_1_TSe	73.3%	0.00	1.1%	R aldehyde
10	RS_3_TS1	1.6%	1.57	RS_C_2_TSe	13.0%	0.67	0.2%	R aldehyde
11	RS_3_TS1	1.6%	1.57	RS_C_3_TSe	9.3%	0.80	0.1%	R aldehyde
12	RS_3_TS1	1.6%	1.57	RS_C_4_TSe	1.9%	1.42	0.0%	R aldehyde
13	RS_3_TS1	1.6%	1.57	RS_H_1_TSe	2.6%	1.29	0.0%	S ketone
14	RS_1_TS1	5.0%	1.12	RS_C_1_TSe	73.3%	0.00	3.7%	R aldehyde
15	RS_1_TS1	5.0%	1.12	RS_C_2_TSe	13.0%	0.67	0.7%	R aldehyde
16	RS_1_TS1	5.0%	1.12	RS_C_3_TSe	9.3%	0.80	0.5%	R aldehyde
17	RS_1_TS1	5.0%	1.12	RS_C_4_TSe	1.9%	1.42	0.1%	R aldehyde
18	RS_1_TS1	5.0%	1.12	RS_H_1_TSe	2.6%	1.29	0.1%	S ketone
19	RS_1_TS1	5.0%	1.12	RS_O_1_TSe	0.0%	5.16	0.0%	RS epoxide

With VRAI:

Table S5.10. Calculation breakdowns for the predicted product percentages of the aldehyde-selective reaction based on VRAI-selectivity calculations only. See Figure 7.C in the main text for elaborations on the VRAI-selectivity calculations. %(Pathway) values are calculated as a product of %(VRAI1) and %(VRAI2). Adding 'jacs13ryu2_d_int2_S' to labels given in TS0, TS1 and TS2 column gives the Gaussian 16 output filename of the relevant structure. Some files may have additional suffixes.

Entry	TS0	TS1	%(VRAI1)	TS2	%(VRAI2)	%(Pathway)	Product
0	R_1_TS0	RR_1_TS1_2	33.90%	RR_C_1_TSe	0.00%	0.00%	S aldehyde
1	R_1_TS0	RR_1_TS1_2	33.90%	RR_H_1_TSe	100.00%	33.90%	R ketone
2	R_1_TS0	RS_1_TS1	66.10%	RS_C_1_TSe	82.20%	54.33%	R aldehyde
3	R_1_TS0	RS_1_TS1	66.10%	RS_H_1_TSe	17.80%	11.77%	S ketone

Table S5.11. Mean absolute error (MAE) calculations: This table presents the mean absolute error (MAE) data of the calculated percentages compared to the experimental percentages. The level of theory for the calculations is ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d).

	MAE compared to the experimental percentage	
	Calculated Percentage (with VRAI)	Calculated Percentage (TST)
Ketone-selective reaction - Cat-B	19.5%	27.0%
Ketone-selective reaction - Cat-C	7.7%	18.5%
Ketone-selective reaction - Cat-D	7.2%	7.3%
Epoxide-selective reaction - Cat-E	5.4%	19.8%
Aldehyde-selective reaction - Cat-F	7.8%	30.7%

5.C Benchmarking

Inclusion of solvent models^{130,174} in single-point energy calculations were considered for key TS1 and TS2 structures with an $\Delta\Delta G^\ddagger < 2.5$ kcal mol⁻¹. The product percentages were re-calculated with the new energy values.

Table S5.12. Benchmarking – inclusion of solvent models: Calculated percentage (TST) results are derived based on the TST assumption. The selectivity of the reaction entirely depends on the kinetics of TS1 and TS2. Calculated percentage (with VRAI) results incorporate the VRAI-selectivity calculation outcomes. It is assumed that the stereochemistry is controlled by kinetics via TS1 and the chemoselectivity, *ie* processes beyond TS1, are controlled by reaction dynamics. Unless otherwise specified, the level of theory for the calculations is ω B97XD/6-311++g(d,p)/SMD/toluene//B3LYP-D3/6-31g(d). The VRAI-selectivity calculations are based on frequency calculation at the B3LYP-D3/6-31g(d) level of theory.

	Experimental Percentage	Calculated Percentage (TST)	Calculated Percentage (with VRAI)
Ketone-selective reaction - Cat-B ^(a)			
Aldehyde	0.00%	45.98%	5.12%
Epoxide	33.30%	1.50%	9.19%
Ketone (<i>R</i>)	44.70%	24.43%	57.62%
Ketone (<i>S</i>)	22.00%	28.08%	28.06%
Ketone-selective reaction - Cat-C ^(a)			
Aldehyde	0.00%	21.10%	14.83%
Epoxide	25.00%	0.00%	23.27%
Ketone (<i>R</i>)	12.00%	0.00%	4.17%
Ketone (<i>S</i>)	63.00%	78.90%	57.73%
Ketone-selective reaction - Cat-D ^(a)			
Aldehyde	0.00%	9.39%	8.25%
Epoxide	11.11%	0.00%	5.60%
Ketone (<i>R</i>)	3.56%	0.00%	2.80%
Ketone (<i>S</i>)	85.33%	90.61%	79.90%
Epoxide-selective reaction - Cat-E ^(b)			
Aldehyde	(40.91%)	69.32%	44.16%
Epoxide (<i>SR</i>)	56.00%	9.93%	43.47%
Epoxide (<i>RS</i>)	0.28%	0.00%	0.00%
Epoxide (<i>SS + RR</i>)	2.81%	20.75%	3.37%
Ketone	0.00%	0.00%	9.00%
Aldehyde-selective reaction - Cat-F			
ω B97XD/6-311++g(d,p)/SMD/toluene //B3LYP-D3/6-31g(d)*			
Aldehyde (<i>S</i>)	1.04%	1.52%	NA
Aldehyde (<i>R</i>)	68.95%	0.00%	NA
Epoxide	0.00%	0.00%	NA
Ketone	30.00%	98.48%	NA
ω B97XD/6-311++g(d,p)/PCM/toluene//B3LYP-D3/6-31g(d)			
Aldehyde (<i>S</i>)	1.04%	0.00%	0.00%
Aldehyde (<i>R</i>)	68.95%	5.14%	54.33%
Epoxide	0.00%	0.00%	0.00%
Ketone	30.00%	94.86%	45.67%

*In the energy check for the VRAI-selectivity test on TS0 as the first transition state, the second transition state (*ie* TS1) is higher in energy by 0.9 kcal mol⁻¹ than the first transition state. In this scenario, the program would proceed to calculate the product percentages with the transition state theory (TST)

Table S5.13. Benchmarking – inclusion of solvent models: This table presents the mean absolute error (MAE) data of the calculated percentages compared to the experimental percentages. Unless otherwise specified, the level of theory for the calculations is ω B97XD/6-311++g(d,p)/SMD/toluene //B3LYP-D3/6-31g(d).

	MAE compared to the experimental percentage	
	Calculated Percentage (with VRAI)	Calculated Percentage (TST)
Ketone-selective reaction - Cat-B	12.1%	26.0%
Ketone-selective reaction - Cat-C	7.4%	18.5%
Ketone-selective reaction - Cat-D	5.0%	7.3%
Epoxide-selective reaction - Cat-E	4.2%	18.5%
Aldehyde-selective reaction - Cat-F	7.8%	32.4%
ω B97XD/6-311++g(d,p)/PCM/toluene //B3LYP-D3/6-31g(d)		

Table S5.14. Benchmarking – inclusion of solvent models – ΔG values related to the dynamically controlled steps: ΔG from TS1 to INT2 ($\Delta G(\text{TS1} \rightarrow \text{INT2})$) and ΔG from INT2 to TS2 ($\Delta G^\ddagger(\text{TS2})$) of the major pathways are recorded below for each reaction. The major pathway is the pathway with the highest %(Pathway). Unless otherwise specified, the level of theory for the calculations is ω B97XD/6-311++g(d,p)/SMD/toluene //B3LYP-D3/6-31g(d). (Unit: kcal mol⁻¹)

	$\Delta G(\text{TS1} \rightarrow \text{INT2})$	$\Delta G^\ddagger(\text{TS2})$
Ketone-selective reaction - Cat-B	-9.7	2.5
Ketone-selective reaction - Cat-C	-10.2	3.3
Ketone-selective reaction - Cat-D	-10.1	3.9
Epoxide-selective reaction - Cat-E	-12.1	0.4
Aldehyde-selective reaction - Cat-F	-5.3	1.3
ω B97XD/6-311++g(d,p)/PCM/toluene //B3LYP-D3/6-31g(d)		

Key TS1, TS2 and INT2 structures with an $\Delta\Delta G^\ddagger < 2.5$ kcal mol⁻¹ at the ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d) theory level were re-optimised at other levels of theory. The product percentages were calculated with the new energy values.

Table S5.15. Benchmarking – structure re-optimisations and product percentages: Calculated percentage (TST) results are derived based on the TST assumption. The selectivity of the reaction entirely depends on the kinetics of TS1 and TS2. Calculated percentage (with VRAI) results incorporate the VRAI-selectivity calculation outcomes. It is assumed that the stereochemistry is controlled by kinetics via TS1 and the chemoselectivity, *ie* processes beyond TS1, are controlled by reaction dynamics.

	Experimental Percentage	Calculated Percentage (TST)	Calculated Percentage (with VRAI)
ω B97XD/6-311g(d,p)// ω B97XD/6-31g(d)			
Aldehyde	0.00%	41.80%	40.60%
Epoxide	33.30%	5.80%	15.90%
Ketone (R)	44.70%	10.30%	10.80%
Ketone (S)	22.00%	42.10%	32.80%
M06-2X/6-311g(d,p)//M06-2X/6-31G(d)			
Aldehyde	0.00%	80.10%	7.40%
Epoxide	33.30%	1.00%	4.80%
Ketone (R)	44.70%	3.70%	59.60%
Ketone (S)	22.00%	15.10%	28.20%
ω b97xd/6-311++g(d,p)//CAM-B3LYP/6-31G(d)			
Aldehyde	0.00%	75.25%	1.86%
Epoxide	33.30%	7.87%	61.57%
Ketone (R)	44.70%	14.26%	35.63%
Ketone (S)	22.00%	2.62%	0.93%

Table S5.16. Benchmarking – structure re-optimisations and errors: This table presents the mean absolute error (MAE) data of the calculated percentages compared to the experimental percentages.

	MAE compared to the experimental percentage	
	Calculated Percentage (with VRAI)	Calculated Percentage (TST)
ω B97XD/6-311g(d,p)// ω B97XD/6-31g(d)	26%	31%
M06-2X/6-311g(d,p)//M06-2X/6-31G(d)	14%	40%
ω b97xd/6-311++g(d,p)//CAM-B3LYP/6-31G(d)	15%	38%

5.D Understand selectivity

The ketone-selective reaction with Cat-D

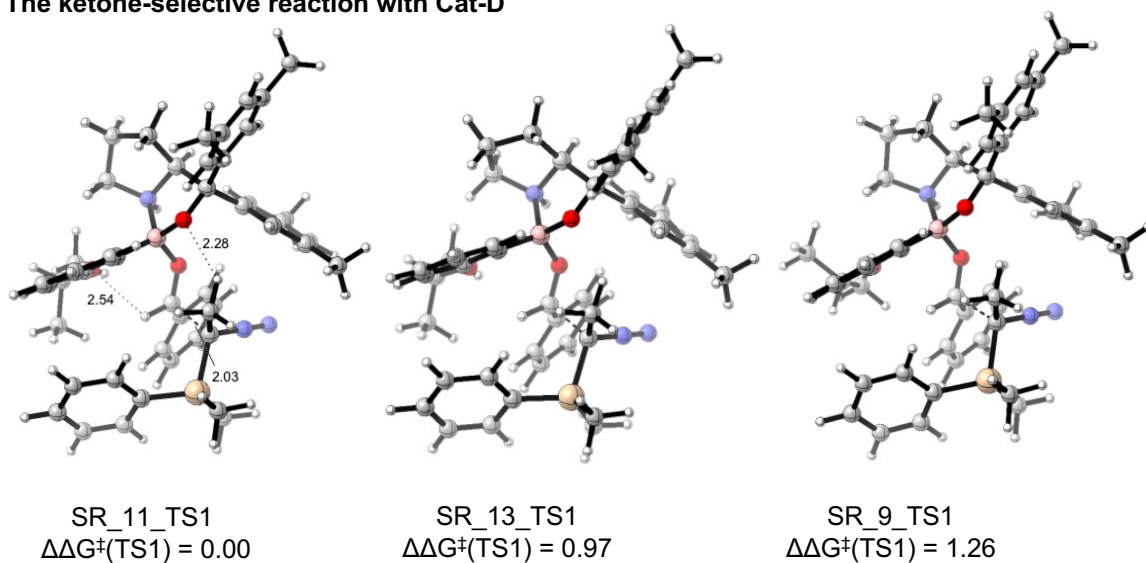


Figure S5.1. Key SR TS1 structures in the ketone-selective reaction with Cat-D. These SR TS1 shares similar geometries and are all within $1.4 \text{ kcal mol}^{-1}$ compared to the global minimum. Level of theory = $\omega\text{B97XD}/6\text{-}311\text{g(d,p)}/\text{B3LYP-D3}/6\text{-}31\text{g(d)}$

Dispersion-interaction analyses:

Table S5.17. Distortion-interaction analysis results. Negative $\Delta U^\ddagger$ were obtained as the association between INT1 and diazo was not accounted, which is likely to be very exothermic. Adding the text in the bracket to the TS1 label in each section gives the Gaussian 16 output filename of the relevant structure. (Unit: kcal mol⁻¹)

TS1	$\Delta U^\ddagger$	$\Delta U_d^\ddagger$	$\Delta U_i^\ddagger$	$\Delta\Delta U_d^\ddagger$	$\Delta\Delta U_i^\ddagger$
(ol17ryu_b_int2_S)	The ketone-selective reaction (Cat-B)				
RR_2_TS1	33.3	89.2	-55.9	6.7	-10.9
RS_1_TS1_2	33.3	88.5	-55.2	6.0	-10.2
SR_3_TS1	35.9	83.0	-47.1	0.5	-2.1
SS_c13_TS1	37.5	82.5	-45.0	0.0	0.0
(ol17ryu_c_int2_S)	The ketone-selective reaction (Cat-C)				
RR_8_TS1	-12.0	32.0	-44.0	3.7	-5.0
RS_5_TS1	-10.6	28.3	-39.0	0.0	0.0
SR_9_TS1	-15.3	38.4	-53.6	10.1	-14.7
SR_e3_TS1	-14.3	39.4	-53.7	11.1	-14.8
SR_e11_TS1	-14.6	39.2	-53.8	10.9	-14.9
SR_e13_TS1	-14.4	39.7	-54.1	11.4	-15.2
SS_13_TS1	-15.8	39.6	-55.3	11.3	-16.4
(ol17ryu_e_int2_S)	The ketone-selective reaction (Cat-D)				
RR_10_TS1	-11.8	32.9	-44.7	7.1	-10.3
RS_9_TS1	-8.6	25.8	-34.4	0.0	0.0
SR_c9_TS1	-12.7	41.7	-54.4	15.9	-20.0
SR_3_TS1	-12.9	42.5	-55.4	16.8	-21.0
SR_11_TS1	-13.9	41.5	-55.4	15.8	-21.0
SR_13_TS1	-13.5	41.8	-55.3	16.0	-20.9
SS_c13_TS1	-14.1	41.8	-55.9	16.0	-21.5
(agw21ryu_d_int2_S)	The epoxide-selective reaction				
RR_2_11_TS1	-16.3	31.9	-48.2	7.4	-4.2
RS_5_TS1	-18.5	33.0	-51.5	8.6	-7.5
SR_2_TS1	-19.0	30.8	-49.8	6.4	-5.9
SR_3_TS1	-18.5	25.5	-43.9	1.0	0.0
SS_10_TS1	-20.2	33.7	-53.8	9.2	-9.9
SS_15_TS1	-20.0	24.7	-44.7	0.3	-0.8
SS_50_TS1	-20.2	24.4	-44.7	0.0	-0.7
SS_65_TS1	-20.3	27.1	-47.4	2.7	-3.5
(jacs13ryu2_d_int2_S)	The aldehyde-selective reaction				
RR_1_TS1_2	-16.6	28.4	-44.9	0.4	-2.9
RR_3_TS1	-14.8	28.8	-43.6	0.8	-1.5
RS_1_TS1	-15.9	28.1	-44.0	0.1	-1.9
RS_2_TS1	-15.9	29.0	-44.9	1.1	-2.8
RS_3_TS1	-15.0	28.0	-43.0	0.0	-1.0
RS_5_TS1	-15.5	28.2	-43.8	0.3	-1.7
SR_2_TS1	-10.9	32.2	-43.1	4.2	-1.0
SS_1_TS1	-11.4	30.7	-42.0	2.7	0.0

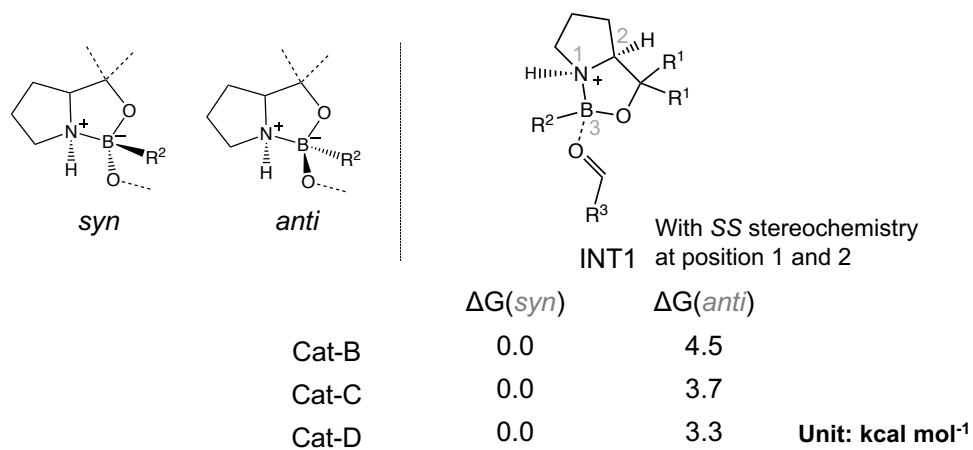


Figure S5.2. The INT1 diastereoisomer stability in ketone-selective reactions: The stereochemistry at position 1 and 2 are SS.

Level of theory = ω B97XD/6-311g(d,p)//B3LYP-D3/6-31g(d)

Appendix to Chapter 6

6.A Data availability

All data and scripts are available in the GitHub repository TwoBondChem/

<https://github.com/Goodman-lab/TwoBondChem>

1. Directory tree of the TwoBondChem/ (as of July 28th 2024)

```
TwoBondChem/
├── bond_strength_descriptor/
│   ├── bond_pre_classifier_v2.py
│   └── bondlength_fromsmi_v2.py
├── data_processing/
│   ├── competitive_pathway_atom_mapping.py
│   ├── initial_treatment.py
│   └── remove_emb_error.py
├── dataset/
│   ├── cyclo_data_v2_16072024.csv
│   ├── da_08012024_vCorr.csv
│   ├── diels_alder_data_v7_19052024.csv
│   ├── exam_test.csv
│   ├── exam_test_28122023.csv
│   ├── fav_RGD1_13012024_v2.csv
│   ├── first-year_data_all_27122023.csv
│   ├── green_04012024.csv
│   ├── green_ea60_c_23052024.csv
│   └── rgd_ea60_c_23052024.csv
├── model_evaluation/
│   ├── chk_dup_v3.py
│   ├── evaluation.py
│   ├── random_sample_test.py
│   ├── random_sample_test_all.py
│   ├── random_sample_test_Bmix.py
│   ├── random_sample_test_Cmix.py
│   ├── save_model.py
│   ├── save_model2.py
│   ├── take_one_out_test.py
│   └── train_size_test.py
├── model_test_2.ipynb
├── React.py
├── README.md
├── RF_model_yr1_28122023.sav
├── scripts/
│   ├── atidx.py
│   ├── bond_classification_01112022.csv
│   ├── get_descriptors_v2.py
│   ├── get_label_v4.py
│   ├── model_training_v3.py
│   └── training_prepare_v5.py
├── cyclo_2b+/
│   ├── cyclo_0_13012024_label.csv
│   ├── cyclo_1_13012024_label.csv
│   └── cyclo_2_13012024_label.csv
```

```
|   |   | cyclo_3_13012024_label.csv
|   |   | cyclo_4_13012024_label.csv
|   |   | cyclo_5_13012024_label.csv
|   |   | cyclo_6_13012024_label.csv
|   |   | cyclo_7_13012024_label.csv
|   |   | cyclo_8_13012024_label.csv
|   |   | cyclo_9_13012024_label.csv
|   | random_sample_test_EXAMPLE.py
|   evaluate_EXAMPLE.ipynb
```

*Archive files have not been included

2. Environment

The scripts in this project are written in Python under the following environment:

- The Python (3.8.12) Standard Library
- Pandas (1.1.5)
- Numpy (1.19.2)
- Sklearn (0.24.2/1.0.1)
- RDkit (2021.9.4)
- Scipy (1.4.1)
- Rxnmapper (0.1.4)

3. Data

dataset/ folder:

All the datasets (Type A: the first-year reactions + three reactions from Part 1A exam at Cambridge; Type B: [3+2] cycloaddition¹⁹⁶ and Diels-Alder reaction dataset;¹⁹⁷ Type C: the Reaction Graph Depth 1 (RGD1))¹⁹⁸ and the Green dataset).¹⁹⁹ are available in the dataset/ folder as csv files. The reaction data is processed and formatted in the same style for the purpose of this investigation. The atom-to-atom mapping numbering in the reaction SMILES is consistent for competitive pathways with the same reactants. There are three columns in each csv file:

- idx: index of the reaction
- code: reactions with the same code are competitive pathways (*ie* having the same reactants)
- reaction: mapped reaction SMILES

The RGD1 dataset csv file also has an 'Rind' column, which corresponds to the index in the 'reaction' column in the original datafile:

https://figshare.com/articles/dataset/model_reaction_database/21066901?file=40272727 (accessed Feb 5th, 2024).

'da_08012024_vCorr.csv' contains 100 Diels-Alder reactions, where the atom-to-atom mapping errors have been picked out and corrected manually.²⁰² Before the correction, reactions with index = 2, 30, 31, 49, 66, 70, 88 and 99 contain errors.

Three out of 147 reactions in the first-year reaction dataset ('first-year_data_all_27122023.csv') had mapping errors, which were subsequently corrected manually. Before the correction, reactions with index = 50, 74 and 124 contain errors. Reactions with a code greater than 77 belong to the testing dataset.

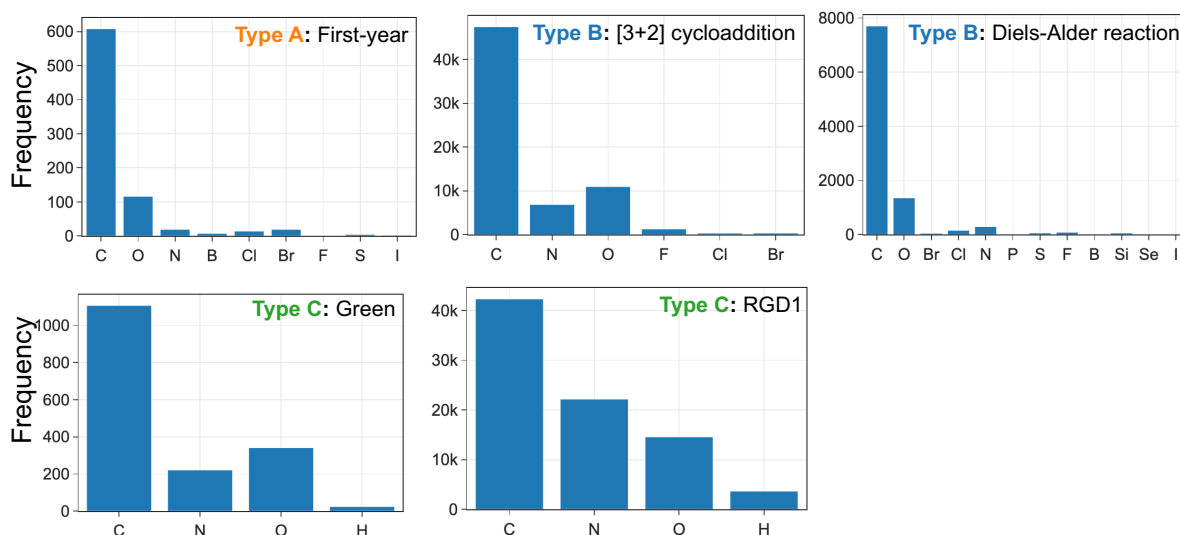


Figure S6.1. An overview of the reactants in each dataset. The bar charts summarise the nature of the non-hydrogen atoms in the molecules.

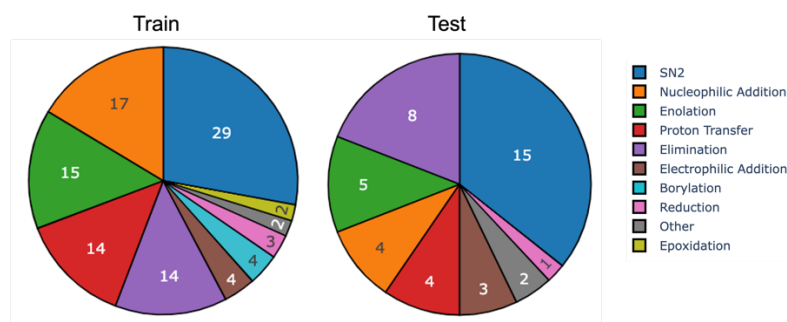


Figure S6.2. Composition of training and testing first-year dataset in model evaluation

4. Scripts

React_v2.py: code for using the random forest model

model_test_3.ipynb: Jupyter notebook on how to use the 'React.py' script

Saved models:

Table S6.1. Saved models in the GitHub

File	Note
RF_model_yr1_28122023.sav	RF model trained from all reactions in the first-year dataset
RF_model_cyclo_11062024.sav	RF model trained from the first 100 reactions in the [3+2] cycloaddition dataset
RF_model_da_11062024.sav	RF model trained from the first 100 reactions in the Diels-Alder dataset
RF_model_green_11062024.sav	RF model trained from the first 100 reactions in the Green dataset
RF_model_rgd_11062024.sav	RF model trained from the first 300 reactions in RGD1 dataset
RF_model_Bmix_11062024.sav	RF model trained from 100 sets of randomly selected [3+2] cycloaddition reactions and 100 sets of randomly selected Diels-Alder reactions.
RF_model_Cmix_11062024.sav	RF model trained from 100 sets of randomly selected Green reactions and 100 sets of randomly selected RGD1 reactions.
RF_model_all_11062024.sav	RF model trained from 78 sets of first-year reactions, 100 sets of [3+2] cycloaddition reactions, 100 sets of Diels-Alder reactions, 100 sets of Green reactions and 100 sets of RGD1 reactions. All reactions were randomly selected from the corresponding full dataset.

scripts/ folder:

atidx.py: contains the functions for conducting atom-to-atom mapping and formatting the reaction SMILES strings for competitive pathways

get_descriptors_v2.py: contains functions for generating the atomistic descriptor components

- Associated file: '**bond_classification_01112022.csv**' – the parameters for generating the bond strength descriptors

get_label_v4.py: contains functions for generating the atomistic label

training_prepare_v5.py: compiles functions in 'get_descriptors_v2.py' and 'get_label_v4.py' to generate the descriptor arrays and labels for atoms in a set of reactants

model_training_v3.py: functions for training and evaluating the model

data_processing/ folder:

initial_treatment.py: group reactions with the same reactants together via InChI strings; for the Diels-Alder dataset only, reactions with placeholder atoms are also removed. Reactions with hypervalent molecules lead to errors in the model training and testing. They are removed manually from the Diels-Alder dataset.

remove_emb_error.py: filter reactions with 3D structure errors

competitive_pathway_atom_mapping.py: This script executes the functions in atidx.py to conduct atom-to-atom mapping of individual reactions and between the competitive pathways. This process was conducted only on the first-year and Diels-Alder datasets.

- Associated files in the dataset/ folder: 'exam_test.csv' and 'exam_test_28122023.csv' – the input and output csv file from executing the script are provided for illustrations

bond_strength_descriptor/ folder:

bondlength_fromsmi_v2.py: gather bond length data from SMILES strings

bond_pre_classifier_v2.py: using the bond length data to generate bond strength descriptor classification criteria, i.e. the 'bond_classification_01112022.csv' file

model_evaluation/ folder:

All the scripts under the model_evaluation/ should be moved from TwoBondChem/model_evaluation/ to TwoBondChem/ before execution.

chk_dup_v3.py: check if there are overlaps between datasets

random_sample_test.py, random_sample_test_all.py, random_sample_test_Bmix.py, random_sample_test_Cmix.py: scripts for carrying out random sampling test

take_one_out_test.py: the script for carrying out the take-one-out cross-validation test

test_size_test.py: The script for investigating the effect of varying the testing or training dataset size

save_model.py: executing this script to train models using reactions from a single dataset and return the trained models in .sav files

save_model2.py: executing this script to train models using reactions from multiple datasets and return the trained models in .sav files

evaluation.py: This script takes the .csv output from the above script to calculate performance metrics by comparing the predictions with the actual labels.

hyperpara_test_v5.py and **para_df.csv**: the script and csv file for conducting hyperparameter tuning

The following files under TwoBondChem/ provide an example of the training and evaluation procedure:

random_sample_test_EXAMPLE.py: training and testing for 2-bond+ models trained from the [3+2] cycloaddition dataset. The code is taken from 'random_sample_test.py' under model_evaluation/ folder. This script can be executed as it is from the TwoBondChem/ folder

cyclo_2b+/ contains output files from executing the 'random_sample_test_EXAMPLE.py' script

evaluate_EXAMPLE.ipynb: evaluation of the predictions using the 'evaluation.py' under model_evaluation/ folder

6.B Bond strength descriptor

Table S6.2. Classification of the bond strength class based on the bond length: bond classes and the associated parameters under each bond type.

Bond Type	Bond strength classes {'bond class name':[minimum length in Å, maximum length in Å]}
CC	{'CC_0': [0, 1.1007], 'CC_1': [1.1007, 1.3593], 'CC_2': [1.3593, 1.4769], 'CC_3': [1.4769, 1.3223], 'CC_4': [1.3223, 1.4429], 'CC_5': [1.4429, 1.4924], 'CC_6': [1.4924, 1.5144], 'CC_7': [1.5144, 1.6075], 'CC_8': [1.6075, 1.784], 'CC_9': [1.784, 3]}
CN	{'CN_0': [0, 1.0569], 'CN_1': [1.0569, 1.3056], 'CN_2': [1.3056, 1.3271], 'CN_3': [1.3271, 1.3602], 'CN_4': [1.3602, 1.233], 'CN_5': [1.233, 1.4057], 'CN_6': [1.4057, 1.4342], 'CN_7': [1.4342, 1.5273], 'CN_8': [1.5273, 1.6554], 'CN_9': [1.6554, 3]}
CH	{'CH_0': [0, 0.9363], 'CH_1': [0.9363, 1.0756], 'CH_2': [1.0756, 1.1072], 'CH_3': [1.1072, 1.2415], 'CH_4': [1.2415, 1.0912], 'CH_5': [1.0912, 3]}
CCI	{'CCI_0': [0, 1.5942], 'CCI_1': [1.5942, 1.7386], 'CCI_2': [1.7386, 1.7251], 'CCI_3': [1.7251, 1.713], 'CCI_4': [1.713, 1.8929], 'CCI_5': [1.8929, 3]}
CO	{'CO_0': [0, 1.1058], 'CO_1': [1.1058, 1.5842], 'CO_2': [1.5842, 1.3065], 'CO_3': [1.3065, 1.4016], 'CO_4': [1.4016, 3]}
HN	{'HN_0': [0, 0.8936], 'HN_1': [0.8936, 1.0002], 'HN_2': [1.0002, 1.0173], 'HN_3': [1.0173, 1.0113], 'HN_4': [1.0113, 1.0108], 'HN_5': [1.0108, 1.0243], 'HN_6': [1.0243, 1.0248], 'HN_7': [1.0248, 1.0293], 'HN_8': [1.0293, 1.1665], 'HN_9': [1.1665, 3]}
CS	{'CS_0': [0, 1.5101], 'CS_1': [1.5101, 1.7539], 'CS_2': [1.7539, 1.6793], 'CS_3': [1.6793, 1.7428], 'CS_4': [1.7428, 1.6302], 'CS_5': [1.6302, 1.8009], 'CS_6': [1.8009, 1.875], 'CS_7': [1.875, 1.9866], 'CS_8': [1.9866, 3]}
OS	{'OS_0': [0, 1.3371], 'OS_1': [1.3371, 1.7609], 'OS_2': [1.7609, 1.4747], 'OS_3': [1.4747, 3]}
CF	{'CF_0': [0, 1.3558], 'CF_1': [1.3558, 1.2366], 'CF_2': [1.2366, 1.3477], 'CF_3': [1.3477, 1.3713], 'CF_4': [1.3713, 1.2977], 'CF_5': [1.2977, 1.4784], 'CF_6': [1.4784, 3]}
NO	{'NO_0': [0, 1.1352], 'NO_1': [1.1352, 1.4542], 'NO_2': [1.4542, 1.5734], 'NO_3': [1.5734, 1.3306], 'NO_4': [1.3306, 3]}
HO	{'HO_0': [0, 0.8564], 'HO_1': [0.8564, 0.9611], 'HO_2': [0.9611, 0.9747], 'HO_3': [0.9747, 0.9752], 'HO_4': [0.9752, 0.9777], 'HO_5': [0.9777, 0.9792], 'HO_6': [0.9792, 0.9797], 'HO_7': [0.9797, 0.9872], 'HO_8': [0.9872, 1.1189], 'HO_9': [1.1189, 3]}
OP	{'OP_0': [0, 1.3884], 'OP_1': [1.3884, 1.7245], 'OP_2': [1.7245, 1.5467], 'OP_3': [1.5467, 3]}

6.C Model evaluation

Algorithm Benchmarking:

Table S6.3. Performance on benchmarking with different machine learning algorithms: random forest (RF), K-nearest neighbour (KNN), support vector (SVC), gaussian process (Gaussian) and multi-layer perceptron classifier (NN). The number of sets of reactions involved in training and testing are specified below in bracket. The '2-bond +' descriptor composition was used.

Model	Accuracy	Precision	Recall
Train: year 1 A (78), Test: year 1 B (30)			
RF	92.04%	86.92%	70.45%
KNeigh	89.49%	80.39%	62.12%
SVC	86.49%	73.33%	50.00%
Gaussian	91.14%	86.87%	65.15%
NN	90.99%	84.62%	66.67%
Train: [3+2] cycloaddition (100), Test: [3+2] cycloaddition (100)			
RF	99.96%	99.77%	100.00%
KNeigh	99.10%	99.75%	94.37%
SVC	97.73%	100.00%	85.21%
Gaussian	99.42%	99.76%	96.48%
NN	99.86%	99.76%	99.30%
Train: RGD1 (300), Test: RGD1 (100)			
RF	81.41%	76.44%	60.69%
KNeigh	80.55%	74.88%	59.16%
SVC	78.90%	74.25%	52.29%
Gaussian	81.28%	77.95%	58.02%
NN	81.10%	76.81%	58.78%

Increasing the size of the training data:

Table S6.4: Model performance when varying the size of the training data. The RGD1 dataset was used for conducting this investigation. The '2-bond +' descriptor composition was used. The unit for training and testing data is 'set', *ie* a set of competitive reactions with the same reactants. The same set of data, 100 sets of RGD1 reactions (*ie* 100 reactions in total), was used for evaluation.

No. of sets of reactions in training	% by atoms			% sets of reactants with:		
	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
100	76.33%	63.10%	62.98%	6.0%	14.0%	28.0%
300	81.22%	76.40%	59.92%	8.0%	30.0%	24.0%
600	81.65%	75.81%	62.79%	6.0%	27.0%	25.0%
900	82.57%	78.80%	62.40%	12.0%	35.0%	27.0%
3000	83.18%	79.57%	63.93%	16.0%	38.0%	25.0%
6000	84.89%	84.54%	64.69%	14.0%	47.0%	26.0%
10300	84.28%	82.75%	63.55%	15.0%	43.0%	28.0%

Performance of models trained with combined dataset:

Table S6.5: Performance of '2-bond +' models trained from the combined dataset. The specific datasets and number of sets of reactions involved are specified below in a consistent format. Let us take the first model presented below, 'Train: 50 (104) [3+2] cycloaddition + 50 (50) Diels-Alder', as an example. 50 sets of [3+2] cycloaddition reactions, equivalent to 104 reactions, with the same reactants and 50 sets of Diels-Alder reactions, equivalent to 50 reactions, with the same reactants are involved in training. In total, there are 154 reactions in the training dataset. Other than the overall results, the performance breakdown by the reactions from different datasets in the testing dataset is also given below.

Testing dataset	% by atoms			% sets of reactants with:		
	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
Type B						
Train: 50 (104) [3+2] cycloaddition + 50 (50) Diels-Alder, Test: 50 (101) [3+2] cycloaddition + 50 (50) Diels-Alder						
All testing data [3+2]	97.55%	91.10%	89.66%	58.0%	85.0%	75.0%
cycloaddition	99.24%	100.00%	95.28%	80.0%	100.0%	80.0%
Diels-Alder	96.52%	83.47%	84.55%	36.0%	70.0%	70.0%
Train: 100 (205) [3+2] cycloaddition + 100 (100) Diels-Alder, Test: 100 (219) [3+2] cycloaddition + 100 (100) Diels-Alder						
All testing data [3+2]	98.23%	94.08%	91.33%	68.0%	88.0%	80.5%
cycloaddition	99.54%	99.76%	97.24%	88.0%	99.0%	89.0%
Diels-Alder	97.37%	88.61%	85.68%	48.0%	77.0%	72.0%
Type C						
Type C: Train: 50 (53) RGD1 + 50 (59) Green, Test: 50 (57) RGD1 + 50 (61) Green						
All testing data	80.34%	76.58%	69.81%	11.0%	31.0%	36.0%
RGD1	79.25%	72.96%	62.27%	4.0%	18.0%	22.0%
Green	81.88%	80.08%	78.14%	18.0%	44.0%	50.0%
Train: 100 (107) RGD1 + 100 (126) Green, Test: 100 (110) RGD1 + 100 (120) Green						
All testing data	79.79%	76.49%	67.30%	20.0%	35.0%	39.0%
RGD1	77.68%	66.00%	62.60%	7.0%	20.0%	31.0%
Green	82.83%	88.73%	72.00%	33.0%	50.0%	47.0%

Reference

- 1 Q. Peng, F. Duarte and R. S. Paton, *Chem. Soc. Rev.*, 2016, **45**, 6093–6107.
- 2 J. P. Reid, I. O. Betinol and Y. Kuang, *Chem. Commun.*, 2023, **59**, 10711–10721.
- 3 M. Elkin and T. R. Newhouse, *Chem. Soc. Rev.*, 2018, **47**, 7830–7844.
- 4 A. J. Sterling, S. Zavitsanou, J. Ford and F. Duarte, *WIREs Comput. Mol. Sci.*, 2021, **11**, e1518.
- 5 H. Hayashi, S. Maeda and T. Mita, *Chem. Sci.*, 2023, **14**, 11601–11616.
- 6 K. B. Lipkowitz and M. A. Peterson, *Chem. Rev.*, 1993, **93**, 2463–2486.
- 7 W. J. Hehre, *A Guide to Molecular Mechanics and Quantum Chemical Calculations*, Wavefunction, Irvine, CA, 2003.
- 8 K. N. Houk, M. N. Paddon-Row, N. G. Rondan, Y.-D. Wu, F. K. Brown, D. C. Spellmeyer, J. T. Metz, Y. Li and R. J. Loncharich, *Science*, 1986, **231**, 1108–1117.
- 9 O. Engkvist, P. O. Norrby, N. Selmi, Y. hong Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discov. Today*, 2018, **23**, 1203–1218.
- 10 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
- 11 L. Goerigk and S. Grimme, *WIREs Comput. Mol. Sci.*, 2014, **4**, 576–600.
- 12 B. L. Foley, M. B. Tessier and R. J. Woods, *WIREs Comput. Mol. Sci.*, 2012, **2**, 652–697.
- 13 N. L. Allinger, *J. Am. Chem. Soc.*, 1977, **99**, 8127–8134.
- 14 N. L. Allinger, Y. H. Yuh and J. H. Lii, *J. Am. Chem. Soc.*, 1989, **111**, 8551–8566.
- 15 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.
- 16 C. Lu, C. Wu, D. Ghoreishi, W. Chen, L. Wang, W. Damm, G. A. Ross, M. K. Dahlgren, E. Russell, C. D. Von Bargen, R. Abel, R. A. Friesner and E. D. Harder, *J. Chem. Theory Comput.*, 2021, **17**, 4291–4300.
- 17 W. Damm, A. Frontera, J. Tirado-Rives and W. L. Jorgensen, *J. Comput. Chem.*, 1997, **18**, 1955–1970.
- 18 V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly and D. L. Mobley, *F1000Research*, **9**, DOI:10.12688/f1000research.27141.1.
- 19 T. Lewis-Atwell, P. A. Townsend and M. N. Grayson, *Tetrahedron*, 2021, **79**, 131865.
- 20 K. Gundertofte, T. Liljefors, P. Norrby and I. Pettersson, *J. Comput. Chem.*, 1996, **17**, 429–449.
- 21 P. H. Y. Cheong, C. Y. Legault, J. M. Um, N. Çelebi-Ölçüm and K. N. Houk, *Chem. Rev.*, 2011, **111**, 5042–5137.
- 22 A. Armstrong, R. A. Boto, P. Dingwall, J. Contreras-García, M. J. Harvey, N. J. Mason and H. S. Rzepa, *Chem. Sci.*, 2014, **5**, 2057–2071.
- 23 M. Besora, A. A. C. Braga, G. Ujaque, F. Maseras and A. Lledós, *Theor. Chem. Acc.*, 2011, **128**, 639–646.
- 24 M. Hajji, N. Abad, M. A. Habib, S. M. H. Elmgirhi and T. Guerfel, *J. Indian Chem. Soc.*, 2021, **98**, 100208.
- 25 S. Bahmanyar, K. N. Houk, H. J. Martin and B. List, *J. Am. Chem. Soc.*, 2003, **125**, 2475–2479.
- 26 B. W. Bakr and C. D. Sherrill, *Phys. Chem. Chem. Phys.*, 2016, **18**, 10297–10308.
- 27 S. Bahmanyar and K. N. Houk, *Org. Lett.*, 2003, **5**, 1249–1251.
- 28 J. P. Reid, L. Simón and J. M. Goodman, *Acc. Chem. Res.*, 2016, **49**, 1029–1041.
- 29 J. C. Golec, E. M. Carter, J. W. Ward, W. G. Whittingham, L. Simón, R. S. Paton and D. J. Dixon, *Angew. Chem. Int. Ed.*, 2020, **59**, 17417–17422.
- 30 S. Gallardo-Fuentes, N. Lezana, S. Lühr, A. Galdámez and M. Vilches-Herrera, *J. Org. Chem.*, 2019, **84**, 10825–10831.
- 31 C. Q. He, C. C. Lam, P. Yu, Z. Song, M. Chen, Y. H. Y. Lam, S. Chen and K. N. Houk, *J. Org. Chem.*, 2020, **85**, 2618–2625.
- 32 A. Mazzanti and D. Casarini, *WIREs Comput. Mol. Sci.*, 2012, **2**, 613–641.

- 33 J. M. Crawford and M. S. Sigman, *Synthesis.*, 2019, **51**, 1021–1036.
- 34 Y. Sohtome and K. Nagasawa, *Chem. Commun.*, 2012, **48**, 7777–7789.
- 35 C. E. Müller, L. Wanka, K. Jewell and P. R. Schreiner, *Angew. Chem. Int. Ed.*, 2008, **47**, 6180–6183.
- 36 M.-Q. Hua, H.-F. Cui, L. Wang, J. Nie and J.-A. Ma, *Angew. Chem. Int. Ed.*, 2010, **49**, 2772–2776.
- 37 D. Y. Rao and A. Anoop, *J. Phys. Chem. C*, 2020, **124**, 582–593.
- 38 J. I. Seeman, *Chem. Rev.*, 1983, **83**, 83–134.
- 39 Schrödinger Release 2021-2 Maestro. Schrödinger LLC, New York, NY, 2021.
- 40 P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *J. Chem. Inf. Model.*, 2010, **50**, 572–584.
- 41 P. C. D. Hawkins and A. Nicholls, *J. Chem. Inf. Model.*, 2012, **52**, 2919–2936.
- 42 RDKit Open-Source Cheminformatics. <https://www.rdkit.org/> (accessed December 18, 2020).
- 43 J.-P. Ebejer, G. M. Morris and C. M. Deane, *J. Chem. Inf. Model.*, 2012, **52**, 1146–1158.
- 44 N.-O. Friedrich, C. de Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey and J. Kirchmair, *J. Chem. Inf. Model.*, 2017, **57**, 2719–2728.
- 45 P. C. D. Hawkins, *J. Chem. Inf. Model.*, 2017, **57**, 1747–1756.
- 46 P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.
- 47 S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, *J. Phys. Chem. A*, 2021, **125**, 4039–4054.
- 48 P. Pracht, S. Grimme, C. Bannwarth, F. Bohle, S. Ehlert, G. Feldmann, J. Gorges, M. Müller, T. Neudecker, C. Plett, S. Spicher, P. Steinbach, P. A. Wesolowski and F. Zeller, *J. Chem. Phys.*, 2024, **160**, 114110.
- 49 S. Axelrod and R. Gómez-Bombarelli, *Sci. Data*, 2022, **9**, 185.
- 50 F. Buttard and P. A. Champagne, *ACS Catal.*, 2022, **12**, 8185–8194.
- 51 M. Bursch, J.-M. Mewes, A. Hansen and S. Grimme, *Angew. Chem. Int. Ed.*, 2022, **61**, e202205735.
- 52 T. Lewis-Atwell, P. A. Townsend and M. N. Grayson, *J. Org. Chem.*, 2022, **87**, 5703–5712.
- 53 A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *WIREs Comput. Mol. Sci.*, 2018, **8**, e1354.
- 54 P. Pechukas, *Annu. Rev. Phys. Chem.*, 1981, **32**, 159–177.
- 55 D. H. Ess, S. E. Wheeler, R. G. Iafe, L. Xu, N. Çelebi-Ölçüm and K. N. Houk, *Angew. Chem. Int. Ed.*, 2008, **47**, 7592–7601.
- 56 P. Bharadwaz, M. Maldonado-Domínguez and M. Srnec, *Chem. Sci.*, 2021, **12**, 12682–12694.
- 57 Z. Yang, X. Dong, Y. Yu, P. Yu, Y. Li, C. Jamieson and K. N. Houk, *J. Am. Chem. Soc.*, 2018, **140**, 3061–3067.
- 58 T. Bekele, C. F. Christian, M. A. Lipton and D. A. Singleton, *J. Am. Chem. Soc.*, 2005, **127**, 9216–9223.
- 59 K. N. Houk, X.-S. Xue, F. Liu, Y. Chen, X. Chen and C. Jamieson, *Isr. J. Chem.*, 2022, **62**, e202100071.
- 60 Y. Oyola and D. A. Singleton, *J. Am. Chem. Soc.*, 2009, **131**, 3130–3131.
- 61 L. M. M. Quijano and D. A. Singleton, *J. Am. Chem. Soc.*, 2011, **133**, 13824–13827.
- 62 K. Jorner, A. Tomberg, C. Bauer, C. Sköld and P.-O. Norrby, *Nat. Rev. Chem.*, 2021, **5**, 240–255.
- 63 S. Lee and J. M. Goodman, *J. Am. Chem. Soc.*, 2020, **142**, 9210–9219.
- 64 S. Lee and J. M. Goodman, *Org. Biomol. Chem.*, 2021, **19**, 3940–3947.
- 65 X.-S. Xue, C. S. Jamieson, M. Garcia-Borràs, X. Dong, Z. Yang and K. N. Houk, *J. Am. Chem. Soc.*, 2019, **141**, 1217–1221.
- 66 P. Yu, T. Q. Chen, Z. Yang, C. Q. He, A. Patel, Y. Lam, C.-Y. Liu and K. N. Houk, *J. Am. Chem. Soc.*, 2017, **139**, 8251–8258.
- 67 S. R. Hare and D. J. Tantillo, *Chem. Sci.*, 2017, **8**, 1442–1449.
- 68 A. Martin-Somer, X.-S. Xue, C. S. Jamieson, Y. Zou and K. N. Houk, *J. Am. Chem. Soc.*, 2023, **145**, 4221–4230.

- 69 D. H. Ess, *Acc. Chem. Res.*, 2021, **54**, 4410–4422.
- 70 R. P. Pemberton and D. J. Tantillo, *Chem. Sci.*, 2014, **5**, 3301–3308.
- 71 Y. J. Hong and D. J. Tantillo, *Nat. Chem.*, 2009, **1**, 384–389.
- 72 K. H. K. Park, N. Frank, F. Duarte and E. A. Anderson, *J. Am. Chem. Soc.*, 2022, **144**, 10017–10024.
- 73 J. H. Hansen, T. M. Gregg, S. R. Ovalles, Y. Lian, J. Autschbach and H. M. L. Davies, *J. Am. Chem. Soc.*, 2011, **133**, 5076–5085.
- 74 C. W. Lee, B. L. H. Taylor, G. P. Petrova, A. Patel, K. Morokuma, K. N. Houk and B. M. Stoltz, *J. Am. Chem. Soc.*, 2019, **141**, 6995–7004.
- 75 E. Martínez-Núñez, *J. Comput. Chem.*, 2015, **36**, 222–234.
- 76 L. P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044–1048.
- 77 P. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043–3050.
- 78 B. Peters, A. Heyden, A. T. Bell and A. Chakraborty, *J. Chem. Phys.*, 2004, **120**, 7877–7886.
- 79 P. M. Zimmerman, *J. Comput. Chem.*, 2015, **36**, 601–611.
- 80 S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu and K. Morokuma, *Chem. Rec.*, 2016, **16**, 2232–2248.
- 81 A. M. Żurański, J. I. Martínez Alvarado, B. J. Shields and A. G. Doyle, *Acc. Chem. Res.*, 2021, **54**, 1856–1865.
- 82 Z. Tu, I. Levin and C. W. Coley, *Enabling Tools and Techniques for Organic Synthesis: A Practical Guide to Experimentation, Automation, and Computation*, 2023, pp. 423–459.
- 83 C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Comput. Phys. Commun.*, 2016, **203**, 212–225.
- 84 T. A. Young, J. J. Silcock, A. J. Sterling and F. Duarte, *Angew. Chem. Int. Ed.*, 2021, **60**, 4266–4274.
- 85 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.
- 86 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
- 87 J. Burés and I. Larrosa, *Nature*, 2023, **613**, 689–695.
- 88 M. A. Kayala and P. Baldi, *J. Chem. Inf. Model.*, 2012, **52**, 2526–2540.
- 89 J. M. Goodman, I. Pletnev, P. Thiessen, E. Bolton and S. R. Heller, *J. Cheminform.*, 2021, **13**, 40.
- 90 P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.
- 91 J. Lu and Y. Zhang, *J. Chem. Inf. Model.*, 2022, **62**, 1376–1387.
- 92 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
- 93 W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, *ACS Cent. Sci.*, 2021, **7**, 1622–1637.
- 94 E. Shim, A. Tewari, T. Cernak and P. M. Zimmerman, *J. Chem. Inf. Model.*, 2023, **63**, 3659–3668.
- 95 M. Fitzner, G. Wuitschik, R. Koller, J.-M. Adam and T. Schindler, *ACS Omega*, 2023, **8**, 3017–3025.
- 96 Z. Tu, T. Stuyver, C. W. Coley, H. Hayashi, S. Maeda and T. Mita, *Chem. Sci.*, 2023, **14**, 226–244.
- 97 F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, *Angew. Chem. Int. Ed.*, 2022, **61**, e202204647.
- 98 M. M. Simpson, C. C. Lam, J. M. Goodman and S. Balasubramanian, *Angew. Chem. Int. Ed.*, 2023, **62**, e202304756.
- 99 A. Breiling and F. Lyko, *Epigenetics Chromatin*, 2015, **8**, 24.
- 100 Y. Lin, X. Yang, W. Pan and Y. Rao, *Org. Lett.*, 2016, **18**, 2304–2307.
- 101 M. D. Hill, *Chem. - A Eur. J.*, 2010, **16**, 12052–12062.
- 102 D. Uredi, D. R. Motati and E. Blake Watkins, *Chem. Commun.*, 2019, **55**, 3270–3273.
- 103 H. Yi, G. Zhang, H. Wang, Z. Huang, J. Wang, A. K. Singh and A. Lei, *Chem. Rev.*, 2017, **117**, 9016–9085.

- 104 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, 2016, Gaussian 16, Gaussian, Inc., Wallin.
- 105 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
- 106 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
- 107 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
- 108 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
- 109 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
- 110 T. Hoshikawa and M. Inoue, *Chem. Sci.*, 2013, **4**, 3118–3123.
- 111 A. McNally, C. K. Prier, D. W. C. Macmillan, M. Andrew, P. C. K. and M. D. W. C., *Science*, 2011, **334**, 1114–1117.
- 112 K. Tanabe, H. Yamada and S. I. Nishimoto, *J. Am. Chem. Soc.*, 2007, **129**, 8034–8040.
- 113 J. D. Cuthbertson and D. W. C. MacMillan, *Nature*, 2015, **519**, 74–77.
- 114 B. Lipp, A. M. Nauth and T. Opatz, *J. Org. Chem.*, 2016, **81**, 6875–6882.
- 115 S. Zhu, J. Qin, F. Wang, H. Li and L. Chu, *Nat. Commun.*, 2019, **10**, 749.
- 116 C. Hansch, A. Leo and R. W. Taft, *Chem. Rev.*, 1991, **91**, 165–195.
- 117 G. Goti, B. Bieszczad, A. Vega-Peñaloza and P. Melchiorre, *Angew. Chem. Int. Ed.*, 2019, **58**, 1213–1217.
- 118 J. J. Zhao, H. H. Zhang, X. Shen and S. Yu, *Org. Lett.*, 2019, **21**, 913–916.
- 119 K. L. Jensen, G. Dickmeiss, H. Jiang, Ł. U. Albrecht and K. A. Jørgensen, *Acc. Chem. Res.*, 2012, **45**, 248–264.
- 120 M. Marigo, T. C. Wabnitz, D. Fielenbach and K. A. Jørgensen, *Angew. Chem. Int. Ed.*, 2005, **44**, 794–797.
- 121 K. S. Halskov, B. S. Donslund, B. M. Paz and K. A. Jørgensen, *Acc. Chem. Res.*, 2016, **49**, 974–986.
- 122 Y. Hayashi, D. Okamura, T. Yamazaki, Y. Ameda, H. Gotoh, S. Tsuzuki, T. Uchimarui and D. Seebach, *Chem. - A Eur. J.*, 2014, **20**, 17077–17088.
- 123 S.-H. Xiang and B. Tan, *Nat. Commun.*, 2020, **11**, 3786.
- 124 E. Arceo, I. D. Jurberg, A. Álvarez-Fernández and P. Melchiorre, *Nat. Chem.*, 2013, **5**, 750–756.
- 125 M. Silvi, C. Verrier, Y. P. Rey, L. Buzzetti and P. Melchiorre, *Nat. Chem.*, 2017, **9**, 868–873.
- 126 E. Le Saux, D. Ma, P. Bonilla, C. M. Holden, D. Lustosa and P. Melchiorre, *Angew. Chem. Int. Ed.*, 2021, **60**, 5357–5362.
- 127 M. Silvi, E. Arceo, I. D. Jurberg, C. Cassani and P. Melchiorre, *J. Am. Chem. Soc.*, 2015, **137**, 6120–6123.
- 128 M. L. J. Wong, A. J. Sterling, J. J. Mousseau, F. Duarte and E. A. Anderson, *Nat. Commun.*, 2021, **12**, 1644.
- 129 A. Vega-Peñaloza, S. Paria, M. Bonchio, L. Dell’Amico and X. Companyó, *ACS Catal.*, 2019, **9**, 6058–6072.
- 130 A. V Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
- 131 P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, *Bioinformatics*, 2009, **25**, 1422–1423.
- 132 F. Pedregosa, V. Gael, A. Gramfort, V. Michel, B. Tririon, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss,

- V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
- 133 G. Luchini, J. V. Alegre-Requena, I. Funes-Ardoiz and R. S. Paton, *F1000Research*, 2020, **9**, DOI: 10.12688/f1000research.22758.1.
- 134 P. Bonilla, Y. P. Rey, C. M. Holden and P. Melchiorre, *Angew. Chem. Int. Ed.*, 2018, **57**, 12819–12823.
- 135 G. Bin Shen, L. Xie, H. Y. Yu, J. Liu, Y. H. Fu and M. Yan, *RSC Adv.*, 2020, **10**, 31425–31434.
- 136 Y. H. Lam, M. N. Grayson, M. C. Holland, A. Simon and K. N. Houk, *Acc. Chem. Res.*, 2016, **49**, 750–762.
- 137 A. Claraz, G. Sahoo, D. Berta, Á. Madarász, I. Pápai and P. M. Pihko, *Angew. Chem. Int. Ed.*, 2016, **55**, 669–673.
- 138 Y. Wu, J. D. Schmitt and R. Car, *J. Chem. Phys.*, 2004, **121**, 1193–1200.
- 139 L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H. B. Li, L. Ding and K. Morokuma, *Chem. Rev.*, 2015, **115**, 5678–5796.
- 140 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
- 141 J. M. Goodman and M. A. Silva, *Tetrahedron Lett.*, 2003, **44**, 8233–8236.
- 142 J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J. P. Piquemal, D. N. Beratan and W. Yang, *J. Chem. Theory Comput.*, 2011, **7**, 625–632.
- 143 A. G. Green, P. Liu, C. A. Merlic and K. N. Houk, *J. Am. Chem. Soc.*, 2014, **136**, 4575–4583.
- 144 E. Larionov, M. M. Mastandrea and M. A. Pericas, *ACS Catal.*, 2017, **7**, 7008–7013.
- 145 R. Aerts, J. Vanhove, W. Herrebout and C. Johannessen, *Chem. Sci.*, 2021, **12**, 5952–5964.
- 146 K. A. Tanemura, S. Das and K. M. Merz, *J. Chem. Inf. Model.*, 2021, **61**, 1647–1656.
- 147 I. Y. Kanal, J. A. Keith and G. R. Hutchison, *Int. J. Quantum Chem.*, 2018, **118**, e25512.
- 148 C. Grondal, M. Jeanty and D. Enders, *Nat. Chem.*, 2010, **2**, 167–178.
- 149 F.-L. Zhang, B. Li, K. N. Houk and Y.-F. Wang, *JACS Au*, 2022, **2**, 1032–1042.
- 150 Z.-L. Li, G.-C. Fang, Q.-S. Gu and X.-Y. Liu, *Chem. Soc. Rev.*, 2020, **49**, 32–48.
- 151 J. M. Smith, S. J. Harwood and P. S. Baran, *Acc. Chem. Res.*, 2018, **51**, 1807–1817.
- 152 Y. Kim and W. Y. Kim, *Bull. Korean Chem. Soc.*, 2015, **36**, 1769–1777.
- 153 Jensen, J. H. xyz2mol, <https://github.com/jensengroup/xyz2mol/blob/master/xyz2mol.py> (accessed May 16, 2024).
- 154 D. G. Nam, S. Y. Shim, H.-M. Jeong and D. H. Ryu, *Angew. Chem. Int. Ed.*, 2021, **60**, 22236–22240.
- 155 L. Gao, B. C. Kang and D. H. Ryu, *J. Am. Chem. Soc.*, 2013, **135**, 14556–14559.
- 156 J. Y. Kim, B. C. Kang and D. H. Ryu, *Org. Lett.*, 2017, **19**, 5936–5939.
- 157 E. J. Corey, R. K. Bakshi and S. Shibata, *J. Am. Chem. Soc.*, 1987, **109**, 5551–5553.
- 158 E. J. Corey, *Angew. Chem. Int. Ed.*, 2009, **48**, 2100–2117.
- 159 J. Y. Kim, Y. S. Lee and D. H. Ryu, *ACS Catal.*, 2021, **11**, 14811–14818.
- 160 S. Y. Shim and D. H. Ryu, *Acc. Chem. Res.*, 2019, **52**, 2349–2360.
- 161 D. H. Ryu and E. J. Corey, *J. Am. Chem. Soc.*, 2004, **126**, 8106–8107.
- 162 L. Zeng, J. Xu, D. Zhang, Z. Yan, G. Cheng, W. Rao and L. Gao, *Angew. Chem. Int. Ed.*, 2020, **59**, 21890–21894.
- 163 B. Thirupathi, S. Breitler, K. Mahender Reddy and E. J. Corey, *J. Am. Chem. Soc.*, 2016, **138**, 10842–10845.
- 164 E. H. E. Farrar and M. N. Grayson, *J. Org. Chem.*, 2022, **87**, 10054–10061.
- 165 R. S. Paton, *Org. Biomol. Chem.*, 2014, **12**, 1717–1720.
- 166 M. N. Paddon-Row, L. C. H. Kwan, A. C. Willis and M. S. Sherburn, *Angew. Chem. Int. Ed.*, 2008, **47**, 7013–7017.
- 167 M. N. Paddon-Row, C. D. Anderson and K. N. Houk, *J. Org. Chem.*, 2009, **74**, 861–868.

- 168 N. Y. M. Omar, N. A. Rahman and S. M. D. Zain, *J. Comput. Chem.*, 2011, **32**, 1813–1823.
- 169 K. Sakata and H. Fujimoto, *J. Org. Chem.*, 2013, **78**, 3095–3103.
- 170 Z. Pi and S. Li, *J. Phys. Chem. A*, 2006, **110**, 9225–9230.
- 171 Y. Wang, X. Guo, M. Tang and D. Wei, *J. Phys. Chem. A*, 2015, **119**, 8422–8431.
- 172 C. C. Lam and J. M. Goodman, *J. Chem. Inf. Model.*, 2023, **63**, 4364–4375.
- 173 C. C. Lam and J. M. Goodman, *Org. Chem. Front.*, 2022, **9**, 3730–3738.
- 174 G. Scalmani and M. J. Frisch, *J. Chem. Phys.*, 2010, **132**, 114110.
- 175 E. E. Kwan and R. Y. Liu, *J. Chem. Theory Comput.*, 2015, **11**, 5083–5089.
- 176 Z. Tu, T. Stuyver and C. W. Coley, *Chem. Sci.*, 2023, **14**, 226–244.
- 177 M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby and O. Wiest, *Org. Lett.*, 2023, **25**, 2945–2947.
- 178 D. S. Wigh, J. M. Goodman and A. A. Lapkin, *WIREs Comput. Mol. Sci.*, 2022, **12**, e1603.
- 179 L. David, A. Thakkar, R. Mercado and O. Engkvist, *J. Cheminform.*, 2020, **12**, 56.
- 180 G. Skoraczynski, P. Dittwald, B. Miasojedow, S. Szymkuc, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.
- 181 S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 1–11.
- 182 S.-Q. Zhang, L.-C. Xu, S.-W. Li, J. C. A. Oliveira, X. Li, L. Ackermann and X. Hong, *Chem. – A Eur. J.*, 2023, **29**, e202202834.
- 183 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.
- 184 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
- 185 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
- 186 S. M. Kearnes, M. R. Maser, M. Wlekinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.
- 187 B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang and G.-W. Wei, *Chem. Rev.*, 2023, **123**, 8736–8780.
- 188 B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos and T. Klucznik, *Chem*, 2018, **4**, 390–398.
- 189 B. A. Grzybowski, T. Badowski, K. Molga and S. Szymkuć, *WIREs Comput. Mol. Sci.*, 2023, **13**, e1630.
- 190 R. Batra, H. D. Tran, C. Kim, J. Chapman, L. Chen, A. Chandrasekaran and R. Ramprasad, *J. Phys. Chem. C*, 2019, **123**, 15859–15866.
- 191 L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton and S. Goedecker, *J. Chem. Phys.*, 2016, **144**, 34203.
- 192 G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler and M. Ceriotti, *J. Chem. Phys.*, 2018, **148**, 241730.
- 193 S. Boobier, Y. Liu, K. Sharma, D. R. J. Hose, A. J. Blacker, N. Kapur and B. N. Nguyen, *J. Chem. Inf. Model.*, 2021, **61**, 4890–4899.
- 194 N. Ree, A. H. Göller and J. H. Jensen, *Digit. Discov.*, 2022, **1**, 108–114.
- 195 J. Clayden, N. Greeves and S. Warren, *Organic Chemistry*, Oxford University Press, 2nd edn., 2012.
- 196 T. Stuyver, K. Jorner and C. W. Coley, *Sci. Data*, 2023, **10**, 66.
- 197 S. Li, X. Wang, Y. Wu, H. Duan and L. Tang, *RSC Adv.*, 2022, **12**, 33801–33807.
- 198 Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev and B. M. Savoie, *Sci. Data*, 2023, **10**, 145.
- 199 C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
- 200 Q. Zhao and B. M. Savoie, *Nat. Comput. Sci.*, 2021, **1**, 479–490.
- 201 Q. Zhao and B. M. Savoie, *Angew. Chem. Int. Ed.*, 2022, **61**, e202210693.

- 202 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2023, **7**, eabe4166.
- 203 A. Lin, N. Dyubankova, T. I. Madzhidov, R. I. Nugmanov, J. Verhoeven, T. R. Gimadiev, V. A. Afonina, Z. Ibragimova, A. Rakhimbekova, P. Sidorov, A. Gedich, R. Suleymanov, R. Mukhametgaleev, J. Wegner, H. Ceulemans and A. Varnek, *Mol. Inform.*, 2022, **41**, 2100138.
- 204 S. Lee, K. Ermanis and J. M. Goodman, *Chem. Sci.*, 2022, **13**, 7204–7214.
- 205 F. Berenger and K. Tsuda, *J. Cheminform.*, 2021, **13**, 88.
- 206 O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.
- 207 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
- 208 C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, John Wiley & Sons Ltd, West Sussex, England, 2nd edn., 2004.
- 209 S. Grimme, A. Hansen, J. G. Brandenburg and C. Bannwarth, *Chem. Rev.*, 2016, **116**, 5105–5154.
- 210 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.
- 211 S. F. Sousa, P. A. Fernandes and M. J. Ramos, *J. Phys. Chem. A*, 2007, **111**, 10439–10452.
- 212 J. P. Perdew, *MRS Bull.*, 2013, **38**, 743–750.
- 213 R. Peverati and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2011, **2**, 2810–2817.
- 214 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.
- 215 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
- 216 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
- 217 Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. <https://plot.ly> (accessed May 16, 2024).
- 218 CYLview20. Université de Sherbrooke. 2020 <http://www.cylview.org/> (accessed May 16, 2024).
- 219 The PyMOL Molecular Graphics System (Version 2.0), Schrödinger LLC., New York, NY, 2015.
- 220 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33–38.