Computational Tools for the Discovery of Novel Aromatic Heterocyclic Bioisosteres

Matthew Theodore Orlando Holland

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2025

Make me to understand the way of thy commandments, and so shall I talk of thy wondrous works.

Psalm 119:27

Computational Tools for the Discovery of Novel Aromatic Heterocyclic Bioisosteres

Matthew Theodore Orlando Holland

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

4th April 2025

Abstract

Aromatic heterocycles are a structural feature of significant importance in small molecule therapeutics; indeed over 85% of the small molecules approved by the FDA between 2020–2024 contain at least one aromatic heterocycle. Key to their utility in medicinal chemistry is their rigid geometric projection of important functionality, the sensitivity with which the physicochemical properties of the parent molecule can be tuned by altering the constitution of the heterocycle, and the range of reliable methodologies enabling their synthesis. Consequently, this structural and physicochemical versatility renders them useful in bioisosteric replacement strategies, where a chemical moiety in a molecule is substituted for another to improve one or more drug-like properties, whilst preserving the biological profile of the initial compound.

Despite the widespread utility of aromatic heterocycles, the proportion of aromatic heterocyclic chemical space that is regularly sampled in medicinal chemistry is limited, and thus constrains the drug-like property space available to drug discovery campaigns. Virtual libraries of compounds enumerated such that they contain large regions of previously unsynthesised molecules have been developed, but tools to explore these for biologically-relevant applications are limited. The work described in this thesis employs the widely-accepted definitions of bioisosterism, relating similarities between molecular shape and electron distribution to broader biological effects, to design and evaluate computational tools that explore regions of aromatic heterocyclic chemical space for new bioisosteres of commonly-occurring heterocycles in medicinal chemistry.

Chapter 1 introduces the relevance of small molecules in drug discovery, and the reasons for the popularity of aromatic heterocycles within these important modalities in medicinal chemistry. An overview of previous work in the field of molecular similarity searching and bioisostere discovery is also presented. **Chapter 2** describes the development and implementation of the first generation of the Heterocycle Isostere Explorer (HCIE), and its merits and limitations are discussed. In **Chapter 3**, the creation of the MoBiVic library of mono- and bifunctionalised aromatic heterocycles is described, leading to the development and implementation of the second generation of HCIE in **Chapter 4**, built around a unique, vector-based alignment algorithm, and a new implementation of electrostatic and shape similarity scoring. The last two chapters describe ongoing experimental work to validate the results of the tools described in the earlier chapters, with **Chapter 5** exploring synthetic strategies for accessing novel aromatic heterocycles, and discussing the challenges encountered in their preparation. **Chapter 6** describes the role of HCIE and other computational tools in ongoing medicinal chemistry projects at the Centre for Medicines Discovery.

Statement of Originality

I, Matthew Theodore Orlando Holland, hereby confirm that this thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images, or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography, or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification.

Signed:

Matthew Holland

4th April 2025

Acknowledgments

In The Clerk's tale, Geoffrey Chaucer writes '*Ay fleeth the tyme; it nyl no man abyde*'; writing these acknowledgements at the end of four years at Oxford I find this sentiment resonating with me as I wonder where the time has gone. The inevitable acceleration of the passage of time is one of two overarching realisations come about through my time here, the other being the power of the extraordinary generosity and kindness that has been extended to me by so many of my friends and colleagues. It is, I fear, impossible in a few pages to acknowledge and thank everyone who has helped and influenced me throughout the past few years, and I apologise in advance to anyone who I have failed to mention here.

I must begin by thanking my remarkable supervisor, Professor Paul Brennan. It was said to me before embarking on a DPhil that the journey could be isolating, scary, and lonely, but the truth is that it has been none of these things and this is in no small part down to you. I cannot think of anyone outside of my immediate family who has had so profound an impact on my life and way of thinking as you have. Your thoughtful and creative scientific guidance, and incredible generosity of spirit fostered within me creativity and independence of thought that I did not think possible, and for that I will forever be grateful. From discussions about pharmacokinetics to lecturing in senate houses in Split to exploding guncotton pigs, I can't think of a single dull moment in your presence. Your mentorship and friendship has guided, supported, developed, and inspired me. I hope that in every stage of my life I am able to show to others the same grace that you have shown to me. Thank you.

The CMD has been a most happy and exciting place to work, and the enjoyment I have derived from my time spent there is down to the wonderful colleagues and friends that I have

been blessed with. Dr Mia Callens has been a source of enormous inspiration to me, alongside Dr Tryfon Zarganis-Tzitzikas, Tom Grimes, Dr Anne-Sophie Marques, Adam Smith, Leonard Lee, Ethan Chidlow, Amber Truepenny, Matt Fry, Amelia Henley, Dr Katherine England, Dr Sheenagh Aiken, Dr Margarida Ruas, Dr Lukas Scheibelberger, Dr Gonzalo Vera Namuncura, Dr Laura Ortega Vargas, Chris Dodd, and everyone at the ODDI and the CMD. I am also grateful to my theory supervisor Prof Fernanda Duarte for her mentorship and support, and to my wonderful colleagues in the Duarte Group, including Dr Tom Young, Aleksy Kwiatkowski, and Martin Flerin. I have had the fortune of working with splendid students throughout my time here, and I am especially grateful to Tom Dyer for certainly teaching me more than I taught him.

My time at Oxford has been greatly improved by the Fellows of Bullingdon College: Dr Nick Elkington, Dr Tristan Johnston-Wood, Felix Drinkall, Tassilo von Müller, and Luke Robinson. Thank you for making the College so wonderful a place to live and exist in, and for putting up with my various foibles and idiosyncrasies with such good humour. I have had such fun with great friends from the various iterations of the DBR, and special thanks must go to Dr Julia Lindsay, Annie Anezakis, Tom Rigney, Saxon Stacey, Tobias Bernard, the Reverend Chris Rimmer, Joshua Bowesman-Jones, Dr Augustin Wambersie, Erin Reelick, Brendan Gliddon, Tobias Schröder, Jack Tottem, and Alex Underwood. I cannot wait for our future adventures together.

The three years I spent singing with the Choir of Merton College definitely include some of my happiest memories, and whilst I am grateful to you all for putting up with me and being a source of such fun, I must mention by name Alexander Smith, Imogen Otley, Amalia Wardle, Joshua Kenney, Kentaro Machida, Benjamin Nicholas, and Eppie Sharp. Dancing with my shirt in tatters on a Harvard rooftop, impromptu discos in the Examination Schools,

and singing at the Animal Ball in Lancaster House in front of the King and Queen will remain with me forever.

Rowing has been a great source of joy, not to mention a useful distraction from science, and although gladly retired from competition I am fortunate to remain involved at various levels. I am grateful for my friends at Cambridge University Boat Club, Oxford University Boat Club, and for the friendship and mentorship that my colleagues on the Boat Race Umpiring Panel have afforded me. Rob Baker has been a steadfast source of guidance and support, for which I am most grateful. I am very proud to be involved in The Sveti Duje International Rowing Regatta, and I am very grateful to my wonderful friends Luka Grubor MBE, Toni Batinić, Toni Gamulin, and Denis Boban for welcoming me into their very special event and making me feel so at home in Croatia. It has been a terrific pleasure working with CD Riches and all involved with the Schools' Head and Marlow Regatta.

I was very fortunate to have been taught at school by some truly inspiring teachers, in particular Dr Richard Kowenicki, Dr Huw Williams, Gilly French, and Dr Hugh Salimbeni, and I am privileged to now count them as friends. At Cambridge I had the immense good fortune to have been mentored by some exceptional scientists, including Sir Alan Fersht FRS, Professor Chris Abell FRS (1957–2020), Dr Andrew Bond, Dr David Summers, Professor Dominic Wright, and Dr Gareth Conduit. The Reverend Dr Cally Hammond and the Very Reverend John Hall have provided me with great comfort and spiritual guidance, and I am grateful to them for this. I will never forget where it all began with the 'Chem Boyos' at Gonville & Caius: Dr Joseph Smith, Dr Fergal Hanna, Dr Matt Haynes, and Thomas Elgar.

Lastly to my family, who have always supported me wholeheartedly in all of my unconventional endeavours. To my mother and father: thank you for your unwavering love and support, and the continual inspiration you provide. To my accomplished and impressive siblings Edward and Lucy: thank you for your support and good humour, and for fostering a terrific atmosphere of fun and a great sense of healthy competition. To my grandparents – Brian, Olive, and Val – thank you for your extraordinary kindness and generosity, and for sparking my curiosity in science and the natural world. To my extended family – Sarah, Philip, Hugo, Carol, Chris, Brenda, and Robert – your love and charity have been unfailing, and I am so deeply grateful for it. To Aloysius, who is always there for me. And finally to Dawn and John, who have stood with me on another shore, and in a greater light: I hope I have made you proud.

Pro donis vestris, gratus ex animo.

Acknowledgments: Funding

The work in this thesis was generously supported by a studentship from the Engineering and Physical Sciences Research Council (grant number EP/S024093/1), for which I am grateful. Exscientia was generous in providing an industrial top-up as part of the SABS R3 Centre for Doctoral Training.

Data Availability

The computer code and data developed and itemised below is freely available on *GitHub*.

i. *HCIE*: `https://github.com/BrennanGroup/HCIE`

ii. *MoBiVic*: `https://github.com/BrennanGroup/MoBiVic`

The computer code for machine learning heterocycle retrosynthesis used in **Chapter 5**, written by Dr Ewa Wieczorek, is also freely available:

iii. *Het-retro*: `https://github.com/duartegroup/Het-retro`

Contents

Abstract i

Statement of Originality iii

Acknowledgments iv

Data Availability ix

List of Figures xv

List of Tables xx

List of Schemes xxii

List of Abbreviations xxiii

1 Introduction 1

1.1 Drugs: Medicine or Poison? 1

1.2 Aromatic Heterocycles in Medicinal Chemistry 9

1.2.1 Hydrogen Bonding 18

1.2.2 Physicochemical Property Optimisation 21

1.2.3 Improving Metabolic Stability 26
1.2.4 Perhaps Flatland isn't so bad 32
1.3 Bioisosterism 35
1.3.1 Bioisosterism: Form Follows Function 35
1.3.2 Bioisosterism in Medicinal Chemistry 38
1.3.3 Aromatic Heterocycles as Bioisosteres 42
1.4 Tools for Identifying Bioisosteric Pairings 43
1.4.1 Pre-existing Datasets 44
1.4.1.1 Mining the PDB and ChEMBL 44
1.4.2 Virtual Screening Approaches 50
1.4.2.1 Virtual Screening Techniques 50
1.4.2.2 Virtual Libraries 54
1.5 Thesis Aims 58

2 Development of the Initial Implementation 60

2.1 Methodology 61
2.1.1 VEHICLe as a Searchable Virtual Library 62
2.1.2 ShaEP as a Searching Algorithm 63
2.1.3 Pre-Processing 66
2.1.3.1 Geometry Optimisation 67
2.1.3.2 Partial Charges 68
2.1.3.3 Mol2 Output File 71
2.1.4 Searching the Database 73
2.1.4.1 Field-Graph Construction 75
2.1.4.2 Subgraph Matching and Transformations 78
2.1.5 Visualisation 80

2.2 Results and Discussion . . . 81
2.2.1 Triazine . . . 81
2.2.2 SARS-CoV-2 Main Protease Inhibitors . . . 84
2.2.2.1 Bioactivity Benchmarking . . . 88
2.2.2.2 Enrichment Factors . . . 92
2.2.3 Methodology Limitations . . . 96
2.3 Conclusions and Future Work . . . 98

3 MoBiVic: The Expansion of the Virtual Library 101

3.1 Introduction . . . 101
3.2 Methodology . . . 104
3.2.1 Substituent Selection . . . 107
3.2.2 Filtering . . . 110
3.3 Results and Discussion . . . 112
3.4 Conclusions . . . 115

4 Development of the Current Implementation 117

4.1 Introduction . . . 117
4.2 Methodology . . . 118
4.2.1 Molecular Searching and Alignment . . . 121
4.2.1.1 Molecule Representation . . . 122
4.2.1.2 Exit-Vector Specification . . . 124
4.2.1.3 Geometry Optimisation . . . 126
4.2.1.4 Partial Charges . . . 131
4.2.1.5 Scoring . . . 137
4.2.1.5.1 Shape Similarity . . . 137

4.2.1.5.2 ESP Similarity 140
4.2.1.6 One-Vector Alignment 145
4.2.1.7 Two-Vector Alignment 149
4.2.1.8 Parallelisation 156
4.2.2 Package Structure 159
4.3 Results and Discussion 160
4.3.1 Bioisosteres for Pyrazolopyridine 160
4.3.2 Rationalising the Activity of Inhibitors of the NLRP3 Inflammasome 166
4.3.3 A Novel Class of Bioisosteres for 2-Pyridine 175
4.3.3.1 Synthesis 180
4.4 Conclusions and Future Work 181
4.4.1 Future Work 184

5 Efforts Towards the Synthesis of Novel VEHICLe Heterocycles 187

5.1 Introduction 187
5.2 Triphosgene Cyclisations 192
5.2.1 Towards the Synthesis of **5.2** 192
5.2.2 Screening of Conditions for the Synthesis of **5.4** 194
5.3 Carbazate cyclisations 195
5.4 Cytosine-Derived Cyclisations 199
5.5 Furoxazinone Synthesis 201
5.6 Conclusions and Future Work 203

6 Translating HCIE to Practice: Experimental Applications in Medicinally Relevant Case Studies 207

6.1 Chemical Probes for NUDT14 . . . 208

6.1.1 Chemical Synthesis . . . 213

6.1.2 Conclusions and Future Work . . . 215

6.2 Small-Molecule Inhibitors of SUV4-20 . . . 217

6.2.1 Chemical Synthesis . . . 224

6.2.1.1 Efforts towards **6.19** and **6.24** . . . 228

6.2.2 Future Work and Conclusions . . . 230

7 Thesis Conclusions 235

8 Experimental Details 241

8.1 Computational Details . . . 241

8.2 General Experimental Details . . . 243

8.3 General Experimental Procedures . . . 246

8.4 Compound Synthesis and Characterisation . . . 248

8.5 Selected NMR Spectra . . . 274

8.6 Biological Assays . . . 283

8.6.1 NUDT5 and NUDT14 Catalytic Assays . . . 283

9 Appendices 284

9.1 Excluding X-H Hydrogens from ESP Similarity Calculations . . . 284

9.2 VEHICLe Sample for Geometry RMSD Benchmarking . . . 289

9.3 Bin Boundaries for Two-Vector Geometry Hashing . . . 290

List of Figures

1.1 Doxorubicin and its side-effects 3

1.2 The diacetylation of morphine and salicylic acid 5

1.3 Small molecules in FDA-approved drugs 10

1.4 Lipinski's Rule of Five 11

1.5 Examples of varying degrees of aromaticity 13

1.6 Some 2023-2024 FDA Approvals 15

1.7 Top aromatic heterocycles in FDA-approved Drugs 17

1.8 Hinge-binding, and Type I kinase inhibitors 20

1.9 Inhibitors of Cathepsin S 21

1.10 Reducing lipophilicity in the development of maraviroc 24

1.11 Oxadiazole isomerisation as a strategy for modulating lipophilicity 25

1.12 The effect of constitutional isomerism of 5,6-bicyclic heterocycles on kinetic solubility 26

1.13 The toxic metabolite of troglitazone 28

1.14 The metabolic soft-spots of indinavir 29

1.15 Increasing the metabolic stability of CB2 agonists 30

1.16 Increasing the metabolic stability of FLT3 inhibitors 31

1.17 Improving the metabolic stability of α7 nAChR agonists 32
1.18 Back to flatland? 34
1.19 Bioisosterism in the development of **deucravactinib** 37
1.20 Bioisosterism to lock ligand conformation 39
1.21 Bioisosterism in FDA-approved drug development 41
1.22 Matched molecular pairs analysis 49
1.23 2D vs 3D virtual screening 51
1.24 The synthetic accessibility of the VEHICLe database 57

2.1 Burger's definition of bioisosterism 61
2.2 Workflow for the first-generation HCIE algorithm 65
2.3 An illustration of the potential energy surface of a molecule 68
2.4 The partial charges of benzene and 4-aminopyridine 70
2.5 An example of a `.mol2` file for pyrazole 72
2.6 A flowchart demonstrating the ShaEP algorithm 74
2.7 The field-graph and molecular shape density surface for triazine 76
2.8 An illustration of a maximal common subgraph 79
2.9 The results of a HCIE search for triazine 82
2.10 The COVID Moonshot pipeline 86
2.11 The chlorobenzylacetamide series of COVID MPro inhibitors 87
2.12 Correlations between ShaEP similarity and pIC_{50} for the chlorobenzylacetamide MPro inhibitors 90
2.13 Enrichment factors for the chlorobenzylacetamide MPro inhibitor series . 93
2.14 A situation where a non vector-based alignment leads to skewed results . 97

3.1 A visualisation of the electrostatic surface potentials of fluorobenzene and pyridine . . . 103
3.2 The algorithm for functionalising the VEHICLe database. . . . 105
3.3 The substituents used to functionalise the VEHICLe database . . . 110
3.4 Examples of potentially explosive or bonkers molecules . . . 111
3.5 A comparison of the medicinally-relevant properties of the VEHICLe, monofunctionalised, and bifunctionalised datasets. . . . 114
3.6 A comparison of the medicinally-relevant properties of the combined datasets with and without PEB molecules. . . . 115

4.1 Molecular representations and atom indexing in HCIE . . . 124
4.2 Molecular geometry optimisation RMSDs . . . 130
4.3 A comparison of the partial charges calculated with various electronic structure methods . . . 133
4.4 Charges by element . . . 134
4.5 Charge comparisons for a random subset of MoBiVic. . . . 136
4.6 The van der Waals molecular volumes for benzene and furan. . . . 138
4.7 Calculating the Tanimoto shape similarity based on volume overlap . . . 139
4.8 The Gaussian approximation to the 1/r curve . . . 142
4.9 The HCIE alignment and scoring algorithm . . . 146
4.10 An illustration of the one-vector alignment and scoring process . . . 147
4.11 Setting up the Kabsch alignment for 2-pyridine. . . . 148
4.12 The parameters used to characterise the geometry between a pair of exit-vectors . . . 150
4.13 The distributions of two-vector parameters for the expanded database of heterocycles. . . . 151

4.14 An illustration of the hashing process 152
4.15 Calculating hashes for thiophene 153
4.16 The structure of the HCIE database 154
4.17 Setting up the Kabsch alignment for a two-vector alignment. 155
4.18 An illustration of the workflow for a two user-vector alignment. 157
4.19 An illustration of the parallelisation process 158
4.20 Flowchart describing the overall alignment and scoring process for a HCIE search. 159
4.21 An illustrative selection of the returned molecules from a HCIE search of 3,5-disubstituted pyrazolpyridine 162
4.22 The alignments of the highest-scoring molecules returned in the HCIE search for 3,5-disubstituted pyrazolopyridine 165
4.23 A matched molecular series where the bioactivity data is too clustered to extract meaningful data 169
4.24 The constant fragments defining the five matched molecular series used in this analysis, and the number of unique ligands present in each series. . . 170
4.25 The correlation plots of the NLRP3 inhibitor matched molecular series . . 172
4.26 The top 10 returned results for 2-pyridine 176
4.27 The proposed 5,5-bicyclic bioisosteres of 2-pyridine in the alignment of highest similarity, and their scores. 179

5.1 The heterocycle retrosynthesis transformer architecture 189
5.2 The VEHICLe heterocycles and their disconnections chosen for synthetic evaluation 191

6.1 Chemical probes for NUDT14 210

6.2 **6.1** in complex with NUDT14 212

6.3 **A-196** and its metabolic profile 218

6.4 Virtual screening hits for SUV4-20 with experimentally confirmed activity 219

6.5 The HCIE search results for 6,7-dichlorophthalazine 234

9.1 The top 10 molecules returned in the HCIE search for **A1** 285

9.2 Isoquinol alignments and atom indexing 286

9.3 The partial charge distributions of isoquinol after alignment 287

List of Tables

1.1 A selection of Langmuir's isosteres 36

3.1 The number of heterocycles in each of the libraries after monofunctionalising the VEHICLe database, and bifunctionalising the monofunctionalised dataset. 107

3.2 The top 10 ring substituents and their frequencies, as extracted from ChEMBL by Hall *et al.* 2017.[290] 109

3.3 A breakdown of the MoBiVic database 112

4.1 Average time taken for molecular geometry optimisation 130

4.2 Average times for single-point calculations 137

4.3 The correlation statistics for the equally-weighted total scores for the NLRP3 MMS. 171

4.4 Optimised weightings and the Pearson correlation coefficients for the total scores 173

4.5 Enrichment factors calculated for 2-pyridine using both HCIE and ElectroShape 177

5.1 Attempted conditions for cyclisation of 3-mercaptothiophene-2-carboxylic acid . . . 194
5.2 Screened conditions for cyclisation of **5.3** . . . 195
5.3 Screened conditions for cyclisation of pyridazine-3-carbaldehyde with alkoxy carbazate . . . 197
5.4 Screened conditions for the cyclisation of isoxazole-3-carbaldehyde with methyl carbazate . . . 198
5.5 Screened conditions for cyclisation of cytosine with 1,1,3,3-tetraethoxy-propane . . . 200
5.6 Screened conditions for the cyclisation of bromo- and phenylcytosine . . . 202

6.1 The selected SUV4-20 compounds and their IC_{50} values . . . 221
6.2 Screen of conditions for S_NAr . . . 226
6.3 Synthesis of **6.16**-**6.18** . . . 227

8.1 Software versions and Python packages used in Chapter 2 and in Chapters 4 and 6. . . . 242

9.1 The RegIDs of the sample of VEHICLe used for geometry RMSD benchmarking. . . . 289
9.2 The distance bins and their hash code. . . . 290
9.3 The angle bins and their hash code. . . . 290

List of Schemes

4.1 The ring-closing metathesis route to quinolizin-4-ones developed by Alanine *et al.* in 2014. 181

4.2 The proposed route to 5,5-heterocycles using ring-closing metathesis. 181

5.1 The synthesis of 3-mercaptothiophene-2-carboxylic acid **5.1**. 193

5.2 The synthesis of bromocytosine **5.16** and phenylcytosine **5.17** from cytosine. 201

5.3 The synthesis of aldehyde **5.9** from commercially available **5.21**. 203

5.4 The attempted cyclisation of **5.9** with hydroxylamine 203

6.1 Retrosynthesis of **6.3** 213

6.2 Synthesis of **6.7** 214

6.3 Synthesis of **6.3** 216

6.4 Retrosynthetic analysis of the SUV420 compound disconnections 224

6.5 Synthesis of tetrachlorophthalazine **6.28** 225

6.6 Rieche formylation and proposed reductive amination of **6.34** 230

6.7 Proposed asymmetric synthesis of **6.20** and **6.25** 232

List of Abbreviations

ADME	Absorption, distribution, metabolism, and excretion
ADMET	Absorption, distribution, metabolism, excretion, and toxicity
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
API	Application programming interface
ATP	Adenosine triphosphate
BSA	Bovine serum albumin
CADD	Computer-aided drug design
CASP	Computer-aided synthesis planning
CDI	Carbonyldiimidazole
CID	PubChem compound identifier
CMD	Centre for Medicines Discovery, University of Oxford
CNS	Central nervous system
COSY	Correlation spectroscopy
COVID	Coronavirus disease
CPU	Central processing unit
CYP	Cytochromes P450
DCM	Dichloromethane

DFT	Density functional theory
DIPEA	*N,N*-Diisopropylethylamine
DMA	Dimethylacetamide
DMF	Dimethylformamide
DMSO	Dimethylsulfoxide
DNA	Deoxyribose nucleic acid
EC_{50}	Half-maximal effective concentration
EF	Enrichment factor
ELS	Evaporative light scattering
ESI	Electrospray ionisation
ESP	Electrostatic surface potential
ETKDG	Experimental-torsion basic knowledge distance geometry
FAM	Carboxyfluorescein
FDA	United States Food and Drug Administration
GI_{50}	Half-maximal growth inhibitory concentration
GPCR	G protein-coupled receptor
HCIE	Heterocycle Isostere Explorer
HIV	Human immunodeficiency virus
HLM	Human liver microsomes
HMBC	Heteronuclear multiple bond correlation
HMT	Histone methyltransferase
HOMO	Highest occupied molecular orbital
HPLC	High-performance liquid chromatography
HRMS	High resolution mass spectrometry
HSQC	Heteronuclear single quantum coherence

HTS	High-throughput screening
IC_{50}	Half-maximal inhibitory concentration
IPA	isopropyl alcohol
IUPAC	International Union of Pure and Applied Chemistry
IV	Intravenous
LCAO	Linear combination of atomic orbitals
LCMS	Liquid chromatography-mass spectrometry
LDA	Lithium diisopropylamine
LRMS	Low resolution mass spectrometry
LRR	Leucine-rich repeat
MB	Megabyte
MCS	Minimal common subgraph
MIT	Massachusetts Institute of Technology
ML	Machine learning
MLM	Mouse liver microsomes
MMFF	Merck molecular force field
MMP	Matched molecular pair
MMS	Matched molecular series
MS	Mass spectrometry
NMR	Nuclear magnetic resonance
PDB	Protein Data Bank
PEB	Potentially explosive, or bonkers
PES	Potential energy surface
pIC_{50}	$-\log_{10} IC_{50}$
QSAR	Quantitative structure-activity relationship

RCM	Ring closing metathesis
RMA	Reactive metabolite assay
RMSD	Root mean square deviation
RT	Room temperature
SAM	*S*-Adenosyl methionine
SAR	Structure-activity relationship
SLSQP	Sequential least squares programming
SM	Starting material
SMARTS	SMILES arbitrary target specification
SMILES	Simplified molecular input line entry system
SVP	Split-valence polarised
$\mathbf{t}_{\frac{1}{2}}$	Half-life
TBAB	Tetrabutylammonium bromide
TCEP	Tris(2-carboxyethyl)phosphine
TEP	Tetraethoxypropane
TFA	Trifluoroacetic acid
THF	Tetrahydrofuran
TLC	Thin layer chromatography
TOCSY	Total correlation spectroscopy
TZVP	Valence triple-zeta polarisation
TZVPP	Triple-zeta valence with two sets of polarisation functions
USD	United States Dollar
USR	Ultrafast shape recognition
UV	Ultraviolet
WHO	World Health Organisation

To my grandparents: Brian Jolliffe, Olive Holland, Dawn Jolliffe (1938 – 2005), and John Holland (1930 – 2021).

1 Introduction

1.1 Drugs: Medicine or Poison?

> οὐ δώσω δὲ οὐδὲ φάρμακον οὐδενὶ αἰτηθεὶς θανάσιμον, οὐδὲ ὑφηγήσομαι συμβουλίην τοιήνδε.
>
> *I will neither give a deadly drug to anybody if asked for it, nor will I make a suggestion to this effect.*

The Hippocratic Oath

Hippocrates (460-370 BC)

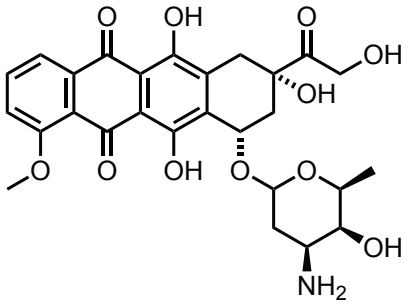
The desire to treat disease, be it to prolong life or alleviate suffering, is one that has driven mankind's interest in understanding the natural world since the earliest points of civilisation. The discovery in 1960 of an almost-complete Neanderthal skeleton in the Shanidar cave

system in northern Iraq was made even more remarkable by the realisation, upon palaeobotanical analysis of the soil surrounding the remains, that the body (known as Shanidar IV) had been intentionally buried on bouquets of eight separate species of flower.[1] Perhaps more striking still is that seven of these plants were known to have been subsequently used extensively in traditional medicines, suggesting that this early civilisation, existing some 65 000 years before the 21st century, had deliberately exploited the natural world for therapeutic benefit.[2,3]

Plants remained the bedrock of what is now known as 'traditional medicine' for the next 63 000 years; a 2001 study reported 122 plant-derived compounds used globally as drugs, suggesting that up to 80% of these owed their discovery to historical ethnomedicine.[4] It is likely, however, that the Neanderthal civilisation's discovery of the medicinal value of these plants was an uncomfortable one driven by trial-and-error consumption and frequented with poisoning and other toxic side-effects.[5]

That substances can be both harmful and therapeutic is well recognised in modern medicine; it is impossible to find an FDA-approved drug that lists no side-effects, and some potent modern medicines (for example chemotherapy agents such as doxorubicin, illustrated in Figure 1.1) have side-effects so severe that their prescription requires careful consideration of whether the therapeutic benefit will outweigh the risks to the patient.[6,7] However, an understanding of the causes of these toxicities and the factors that govern the balance between harmful and helpful were not well understood until the beginning of the 19th century.[8]

The Ancient Greek physician Hippocrates is widely regarded as being the first to bring medicine out of the grip of theology and into the realms of science by advocating for a systematic approach to medicine and pharmacology based on observation, a reputation which



Doxorubicin

Side effects

Very common
- Infection
- Lack of apetite
- Vomiting and nausea
- Fever
- Hair loss
- Liver abnormalities

Common
- Blood poisoning
- Long-term heart problems
- Conjunctivitis
- Heart failure
- Inflammation of the throat and gullet

Figure 1.1. Medicine or poison? The structure of the chemotherapeutic agent doxorubucin, and its very common (affecting more than 1 in 10 patients) and common (affecting 1 in 10 patients) side effects, as outlined in the patient information leaflet.

widely affords him the moniker '*The Father of Western Medicine*'.[9] One of many crucial insights was that diseases had natural causes, rather than supernatural or divine ones, and that natural therapies (for example those based on herbs and plants) could be used to correct the imbalances causing these illnesses. Hippocrates also introduced the concept that treatments or drugs administered must be guided by the nature of the symptoms and the patient, hereby linking individual illnesses (described in this case by sets of phenotypic symptoms) to specific therapies and producing an early attempt at balancing the drug with the toxin; an early recognition that selecting the appropriate treatment for the symptoms presented can favour the therapeutic benefits over harmful side-effects.[10–12]

Hippocrates' influence on medicine, and what later came to be known as drug discovery, is significant; not only is his oath still sworn by physicians today, but his introduction of a rudimentary scientific method to the treatment of human disease has arguably contributed greatly to the discovery of many biologically active substances.[13] A later Greek physician Galen (131-201 AD) took on the task of extending and widely documenting the work that Hippocrates had started, and was key in introducing (amongst a great many other things)

the concept of polypharmacy, and first recorded the anti-inflammatory effects of willow bark (now known to contain salicylic acid, a precursor to aspirin).[9,14,15]

The next greatest influence on modern medicinal chemistry arguably came from Philippus Aureolus Theophrastus Bombastus von Hohenheim (1493 – 1541; a name so unwieldy he is commonly known as Paracelsus), a Swiss chemist and physician whose work laid the foundations of modern toxicology, and introduced for the first time chemical procedures for the production of medically active concoctions (principally from laudanum,[a] tartar,[b] and vitriol.[c])[9,16–18] In his Third Defence (published posthumously in 1564), Paracelsus famously asserted that '*Solely the dose determines that a thing is not a poison*', hereby introducing for the first time formally a notion of an observed effect depending on the quantity administered, and also advocated for the use of inorganic compounds of metals as treatments for disease.[19] Where Hippocrates and Galen had provided ideas about the specificities of treatments to individual diseases and provided a rational and experimental framework for treating disease, Paracelsus now introduced a link between chemistry (and by association, chemical compounds) and medicines, and went some way towards providing a link between the amounts (what we would now know as concentration) of a substance in the body and the biological effect.

The 19$^{\text{th}}$ century saw several great scientific revolutions, and the pace of discovery increased dramatically. Of significant importance in the foundations of the field that became medicinal chemistry was the extraction of morphine from '*poppy juice*' by the German pharmacist Friederich Wilhelm Sertürner in 1805, demonstrating the importance of isolating active compounds from biological extracts for targeted therapeutic effects, and ultimately leading to

[a] A tincture of opium now known to contain morphine.

[b] Potassium bitartrate

[c] Sulfuric acid

the synthesis by Heinrich Drayer, and subsequent production and marketing in 1895 by the German drug company Bayer, of morphine's diacetylated cousin heroin.[20–22] The significance of this discovery was a realisation that the observed biological effect of a tincture or concoction could be attributed to one (or more) active ingredients, and crucially that this active ingredient was crystalline in its pure form (a chemical compound). These chemical transformations are summarised in Figure 1.2.

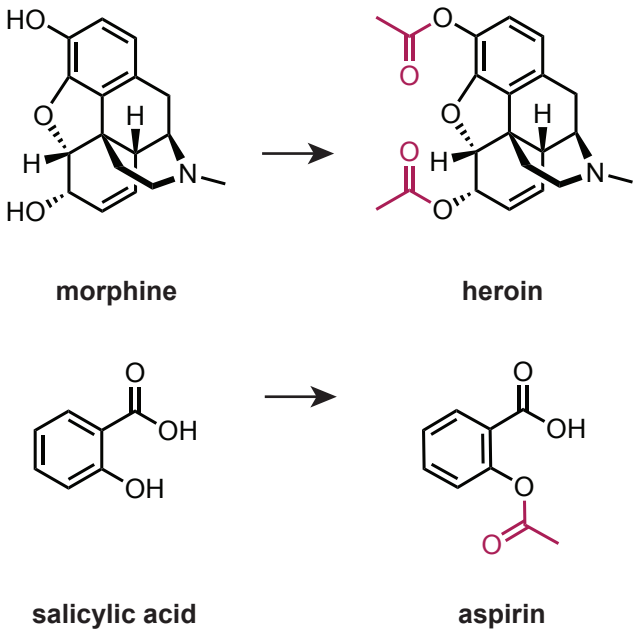


Figure 1.2. The structures of morphine and salicylic acid, and their acetylated products. The diacetylation of morphine at Bayer intended to create a less potent and addictive form of the analgesic; history has not judged this transformation kindly. Aspirin was created from salicylic acid, a popular ingredient in traditional anti-inflammatory medicines originally sourced from willow bark, to reduce propensity of the latter to induce stomach irritation.

The synthesis of acetylsalicylic acid (aspirin) by Arthur Eichengrün and Felix Hoffmann (both working at Bayer) in 1897 saw firmly established the link between the newly developing field of synthetic organic chemistry and medicine discovery;[d] a relationship between the chemical modification of naturally-derived products and their biological properties had been clearly

[d]There is some controversy as to whom the discovery of aspirin should actually be attributed. Discussion of this is outside the scope of this brief introduction, but for an excellent account see Sneader (2000).[23]

demonstrated.[22–24] Chemical synthesis now provided a tool by which the potency and side-effects of medicinal compounds could be altered, bringing 20th century medicinal chemists closer to controlling the balance between poison and medicine precisely and rigorously.

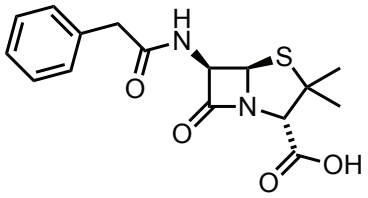
The 'germ-theory' of infection and disease formally introduced by the French chemist Louis Pasteur in 1878 laid the foundations for what is now termed molecular medicine, and is rightfully seen as a discovery of enormous significance in the journey towards specific and selective medicines.[25–27] The discovery of these causative agents of infection, and laboratory techniques to isolate, grow, and observe them, paved the way for the first example of phenotypic screening by the German Paul Ehrlich in 1909, leading to the development of the first antibiotic: the arsenic-based agent arsphenamine as an effective treatment for syphilis[e] (marketed in 1910 by the German firm Hoechst AG as Salvarsan).[27] Whilst the importance of this discovery, the first ever antibiotic, is evident, it is somewhat overshadowed in history by Sir Alexander Fleming's identification in 1928 of penicillin, and its subsequent purification by Howard Florey at Oxford, for which they shared the 1945 Nobel Prize in Physiology or Medicine.[27,29] The publication of the X-ray structure of penicillin by Dorothy Hodgkin in 1949 (confirming unambiguously the presence of a β-lactam ring) enabled the production of semi-synthetic analogues that allowed the properties of these molecules, including their propensities towards the development of resistance, to be adjusted with atomic precision, hereby kickstarting the 'Golden Age' of antibiotic discovery from the 1940s-1960s.[30] Now medicines could be adjusted precisely at the molecular level, guided by molecular structures determined by X-ray, and the effect of these atomic changes on their biological properties began to be observed and rationalised.

[e]This discovery was, surprisingly, not universally welcomed and Ehrlich was subject to abuse both from those who feared that treating sexually-transmitted disease would lead to a breakdown in societal morality, and those with antisemitic motivations.[28]

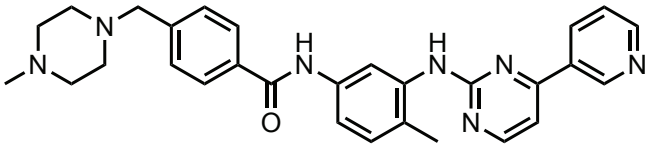
The rapid advancements in the understanding of disease biology and the determining of molecular structures in the latter half of the 20th century saw the balance tip further from poison to drug. The new field of pharmacokinetics was able to rationalise why certain molecular structures caused harmful or off-target effects, and the introduction of high-throughput screening (HTS) in 1989 enabled large libraries of molecules to be searched comparatively quickly for new drugs.[31,32] Improvements in synthetic organic methodologies enabled the design and construction of elaborate chemical structures, and equipped medicinal chemists with the tools to control carefully the architectures of their new drugs to maximise selectivity and reduce side effects and toxicity.[33] Rapid developments in protein structure determination, especially by NMR spectroscopy and X-ray crystallography, made structure-based rational drug design a possibility, with the first molecules discovered in this manner (both HIV protease inhibitors) published in 1990.[34–36]

These improvements in understanding, driven by technology, allowed medicinal chemists to become more ambitious in their attempts to drug challenging diseases and targets. The FDA-approval in 2001 of imatinib, the first kinase inhibitor (prior to this kinases were considered to be undruggable targets, as the conserved ATP-binding pocket across so large a family of enzymes raised concerns about selectivity) was a remarkable achievement, changing the average five-year survival of patients with chronic myeloid leukemia from 31% in 1993 to over 90% in 2023, and has since lead to the FDA-approval of over 70 further protein kinase inhibitors.[37–39]

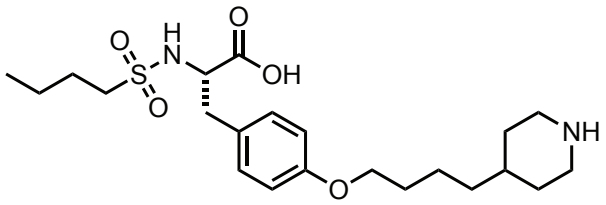
This period also saw a revolution in computer technology, and with it the realisation that *in silico* techniques could improve both efficiency and sustainability in drug discovery.[40] A paper published in Science in 1980 by scientists working at the German pharmaceutical firm Merck outlined a rudimentary virtual molecular modelling method used in several of their



penicillin



imatinib



tirofiban

drug discovery campaigns.[41] Papers on protein-ligand docking soon followed in 1982, and throughout the 1980s significant investment in the new field of computer-aided drug discovery (CADD) was made by pharmaceutical firms, eventually leading to the FDA approval in 1999 of tirofiban, a fibrinogen receptor antagonist indicated for the prevention of blood clots during heart attack and surgery, whose lead was discovered in a pharmacophore-based virtual screen.[42–44] In 2010, 12 small molecule drugs were either approved or in clinical trials that had been discovered or optimised using CADD, and although more recent reviews are lacking, this number is likely to be increasing.[45]

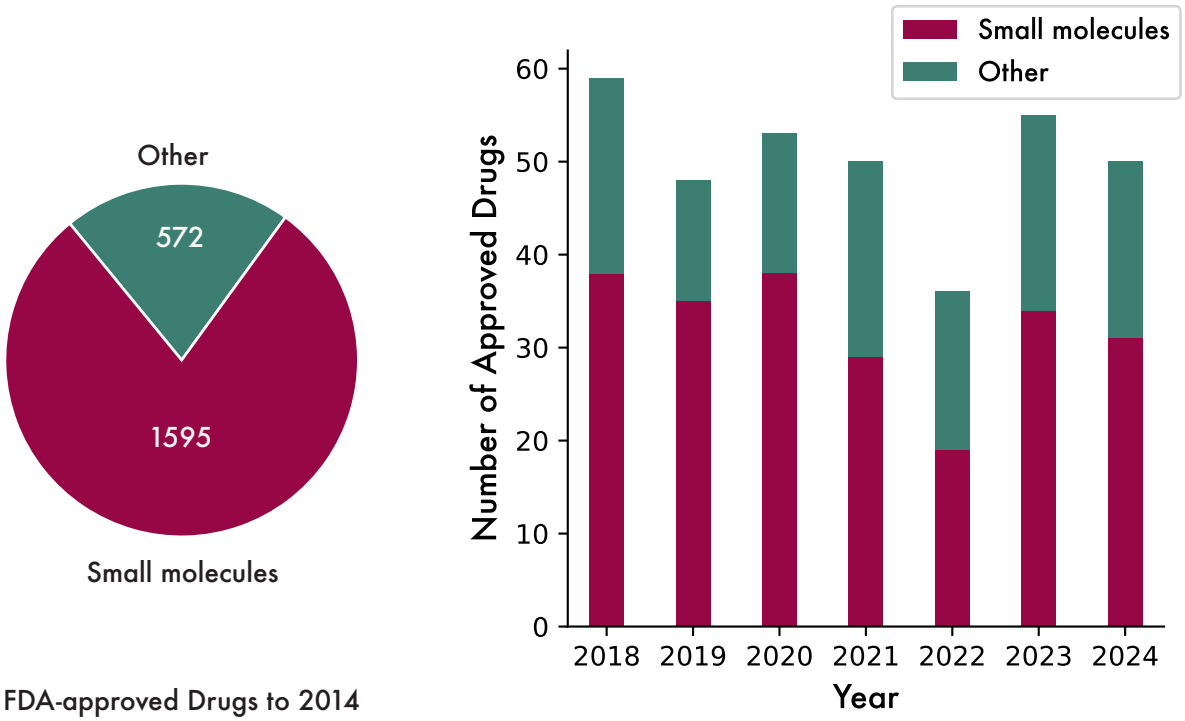
While the Greek physicians laid the groundwork for the scientific method, Paracelsus integrated chemistry into medicine, the Bayer chemists of the 1890s pioneered synthetic modification, and the antibiotic revolutionaries of the early 20th century advanced the understanding of chemical structures and specific targets, computers were now beginning to introduce another paradigm to drug discovery. The ability to search rapidly and methodically through large virtual libraries of compounds, and to learn from large volumes of data using machine-learning algorithms developed in physics has become more readily accessible to medicinal

chemists as the cost of computing power decreased, and the computer algorithms became more sophisticated.[46]

It is estimated that the cost of developing a drug from inception to approval is of the order of $2 billion USD, taking an average of 15 years.[47] A large proportion of this cost is due to high attrition rates of molecules in the clinic, with an average of 90% of clinical trials failing to meet their objectives, with failure most frequently attributed to a lack of clinical efficacy, poor toxicology, or poor exposure due to unfavourable drug-like properties.[48] The integration of computational methodology into drug discovery pipelines has the potential to significantly reduce this attrition rate by aiding the selection of higher-quality leads with better drug-like and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, further shifting the dial from poison to medicine.[46] The decision by the Royal Swedish Academy of Sciences to award the 2024 Nobel Prize in Chemistry to David Baker of the University of Washington, USA for his work on computational protein design, and Sir Demis Hassabis and John Jumper, both of Google DeepMind in the United Kingdom, for their contributions towards protein structure prediction, serves to highlight the potential that computational techniques have to improve mankind's ability to treat disease.[49]

1.2 Aromatic Heterocycles in Medicinal Chemistry

Despite the gradually increasing prevalence of biologics amongst recent FDA-approvals, small molecules still account for the majority of drug approvals in the years 2018–2024 (Figure 1.3b), and with a 2014 analysis of all FDA-approved pharmaceuticals finding over 75% of these to be small molecules (Figure 1.3a), it is clear that this class of therapeutics is of significant importance in medicinal chemistry.[50–58]



(a) The proportion of all FDA-approved drugs up to 2014 that are small molecules (1595 drugs). Other includes biologics (146), peptides (23), and combination drugs (253).[58]

(b) The number of drugs approved by the FDA annually between 2018 and 2024, showing the relative proportions of small molecules.[51–57]

Figure 1.3. Small molecules in FDA-approved drugs.

The medicinal appeal of small molecules is linked to the ease with which their physicochemical properties and capacity for specific intermolecular interactions, which govern their pharmacokinetics and pharmacodynamics, can be tuned.[59,60] The factors that govern the ADMET profiles of small molecules are well-understood at an atomic resolution.[61–63] As success in clinical trials has been shown to be highly correlated with an understanding of ADMET properties at the discovery and lead stage, the advantage that small molecules provide in this regard is significant, hence their broad appeal in modern drug discovery.[64,65]

A now classic example of atomic-level understanding of ADMET properties and the development benefits these afford is given by Christopher Lipinksi's pioneering analysis of orally-

bioavailable FDA-approved drugs.[66] Oral delivery is the preferred route of administration for the majority of pharmaceutical products, due to its non-invasive simplicity, convenience leading to high patient compliance, and cost-effectiveness in large-scale manufacturing.[67] However, the success of an orally administered drug depends on its ability to dissolve and be absorbed in the gastrointestinal tract, withstand first-pass metabolism, and reach systemic circulation at therapeutically relevant concentrations. As a result, a molecule's physicochemical properties, such as solubility and metabolic stability, play a crucial role in determining its oral bioavailability (the fraction of the orally administered drug that reaches systemic circulation). Lipinski proposed a set of guidelines, now immortalised as the 'Rule of Five', which offer a useful approximation of a drug's likely oral bioavailability (see Figure 1.4).[66] Subsequent analyses have since shown that 'Rule of five' compliant molecules have a higher success in clinical trials, and of gaining FDA approval.[68,69]

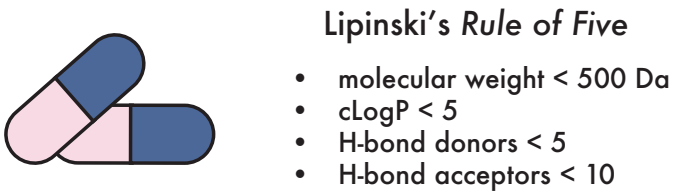


Figure 1.4. The 'Rule of Five' for orally-bioavailable small molecules, as proposed by Christopher Lipinski.[66]

Alongside this raised probability of clinical success, small molecules are (for a large part) reliably accessible through synthetic organic chemistry, and their characterisation and purification is dependable and well-understood.[60,70] Their modular nature, where distinct molecular regions mediate specific target interactions, combined with modern synthetic strategies allows for efficient structural optimisation of lead compounds. This flexibility facilitates the fine-tuning of key properties such as target binding affinity, selectivity, and metabolic stability.[60,71]

Featured widely as structural elements across small molecule therapeutics are aromatic heterocycles. Although the exact definition of aromaticity is a controversial topic amongst chemists, for the purposes of this work aromatic heterocycles are defined as cyclic molecules of carbon or other heteroatoms[f] fulfilling the following criteria:

a. that each ring atom is able to participate in π bonding (the ring is fully conjugated);

b. the number of π electrons must be $4n + 2$, $n \in \mathbb{Z}$ (often referred to as 'Hückel's rule'); and

c. the ring(s) must be flat.[72,73]

In carbocyclic systems, the distinction between aromatic ($4n + 2$ π electrons), anti-aromatic ($4n$ π electrons), and non-aromatic is typically well-defined. In contrast, heterocyclic compounds exhibit a far broader range of aromatic character, with significant variation in the degree to which aromaticity is expressed (Figure 1.5).[74] This reduced aromaticity (when compared to benzene) arises from poorer π-delocalisation of the heteroatomic electrons, due to either poorer overlap with the carbon p orbitals (in the case of sulfur with its diffuse 3p orbitals) or the tendency of highly electronegative elements like oxygen to localise their lone pairs, reducing their contribution to the extended π system.

The effect of this reduced aromatic character is illustrated in Figure 1.5, where physical characteristics indicative of aromaticity are compared for a series of simple heterocycles. Benzene, which is taken as the benchmark of complete aromaticity in this analysis, exhibits full bond length equalisation reflecting complete delocalisation of the carbon 2p electrons into a fully extended π system. The significantly downfield chemical shift of the benzene protons

[f] Benzene is therefore included within this definition.

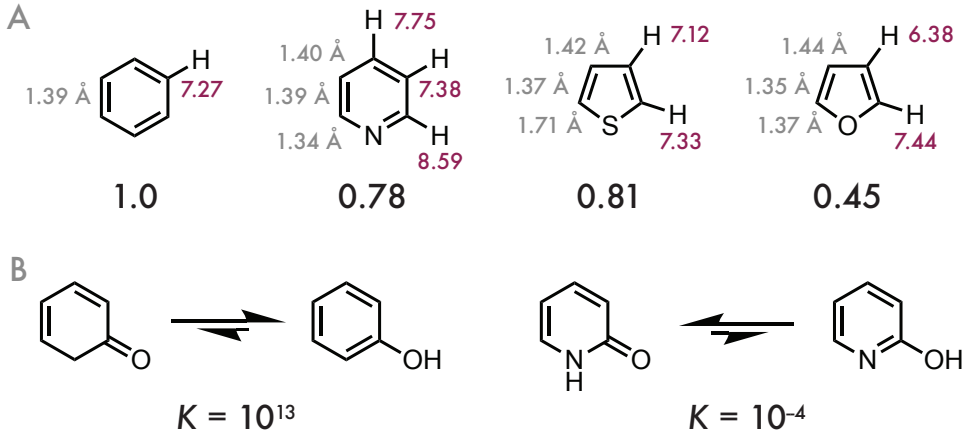


Figure 1.5. The varying levels of aromaticity exhibited by heterocyclic compounds. **A** | The bond lengths (gray), proton NMR spectroscopic chemical shifts (ppm; red), and relative aromatic resonance energies (black) of benzene, pyridine, thiophene, and furan.[75] **B** | The tautomers of phenol and 2-pyridone, and their equilibrium constants in aqueous solution.[76,77]

(δ 7.27 ppm) compared to those of non-aromatic cyclohexene (δ 5.66 ppm) indicate the induction of a diatropic ring current by the external magnetic field of the NMR spectrometer; again a hallmark of a fully delocalised π system.

In contrast, examination of bond length patterns in pyridine, thiophene, and furan reveals varying degrees of alternation, suggesting a more limited contribution of the heteroatom to the delocalised π system. This is reflected in the relative aromatic resonance energies[g], all of which are lower than that of benzene. Interestingly, although sulfur's diffuse 3p orbitals are generally considered to overlap less effectively with the carbon 2p orbitals, thiophene's resonance energy is slightly higher than pyridine's; it has been (controversially) suggested that this may result from some involvement of sulfur's 3d orbitals in the π-system.[78]

The proton chemical shifts of these heterocycles also support this trend, albeit with heteroatomic inductive withdrawal causing downfield shifts of the α-protons. That these chem-

[g]The difference in calculated energy between the ground state energy of the heterocycle and that of the equivalent, hypothetical non-aromatic cyclic polyene.

ical shifts vary intramolecularly is indicative of the uneven distribution of electron density about the π system. Nonetheless, all protons resonate at more downfield values than would be expected for their semi-saturated equivalents, suggesting the presence of a diatropic ring current. This suggests that, despite reduced aromatic character, these heterocycles still display sufficient π-delocalisation to be reasonably classified as aromatic.

The level of aromaticity also impacts the dominant tautomeric form in aqueous solution, as illustrated in Figure 1.5 panel B. Phenol's keto tautomer results in a saturated carbon α to the carbonyl and a loss of ring aromaticity, whereas the enol form retains the aromatic ring, thus the enol form is very significantly favoured. For 2-pyridone both the lactim and the lactam forms can be considered to be aromatic (the lone pair on nitrogen in the lactam is considered to be delocalised into the ring), and thus the preference for one form over another is much weaker. The moderate preference for the lactam over the lactim in this case can be attributed to both the greater bond strength of the carbonyl over the imine, and the greater polarity of the lactam interacting more strongly with the polar solvent.

Inspection of the small molecule FDA approvals in 2023 and 2024 reveal that 59 out of 64 of these contain aromatic heterocycles, and they are represented in molecules across all indications.[56,57] Figure 1.6 shows a selection of these molecules across a range of indications. It is clear even from this small sample that heterocycles in drugs include a diverse range of heteroatoms, ring sizes, and ring fusions.

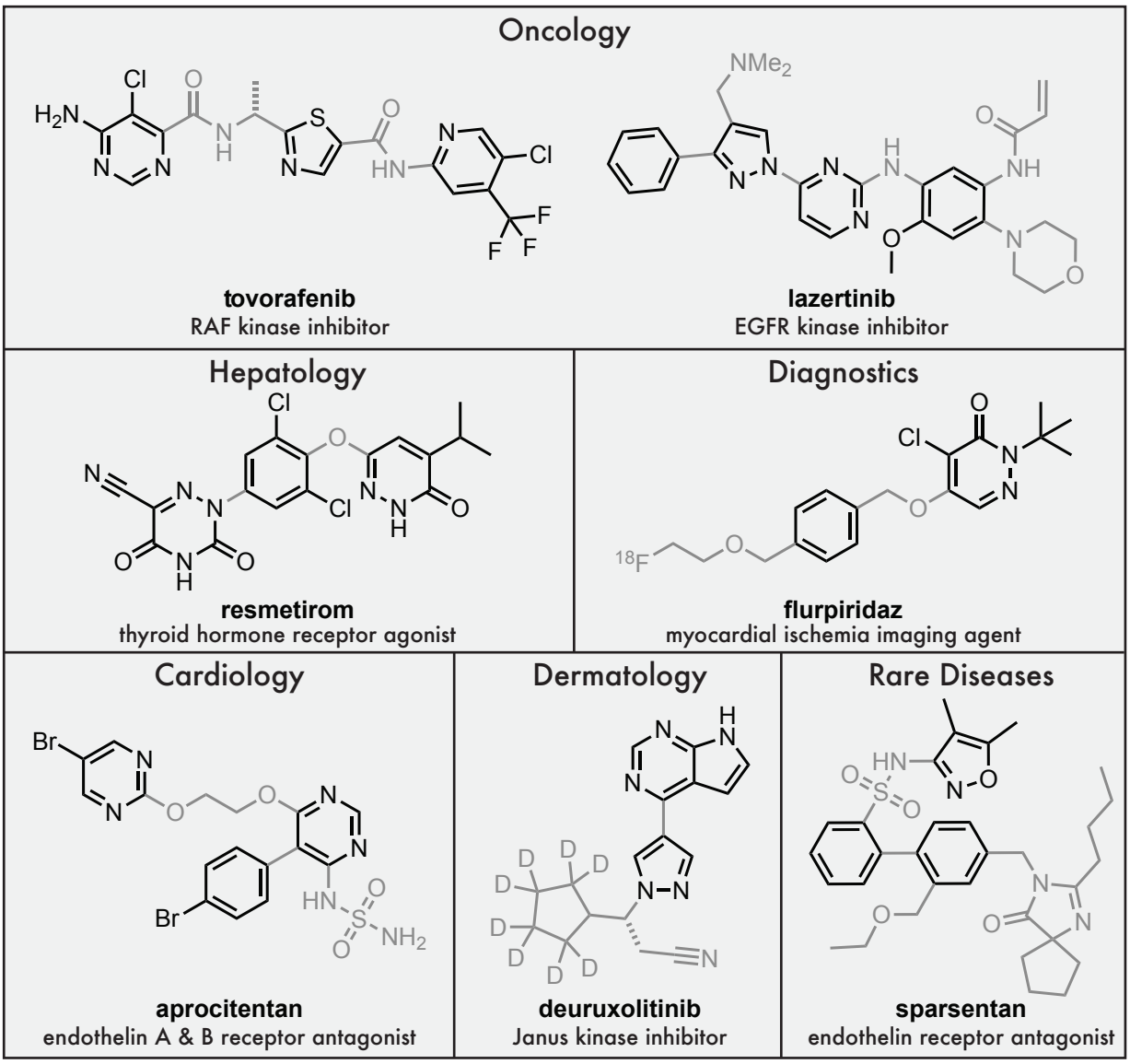


Figure 1.6. A selection of drugs approved by the FDA in 2023 or 2024 and grouped by indication. Each drug contains at least one aromatic heterocycle, which are highlighted in bold.

This pattern is reflected in the body of FDA-approved drugs as a whole. An analysis of all FDA-approved drugs up to December 2012 by Vitaku *et al.* found that 640 (59%) of the 1086 unique small molecule drugs approved by the FDA since 1938 contained nitrogen heterocycles (this figure includes aliphatic heterocycles like piperidine), with aromatic heterocycles making up 12 of the top 25 most frequent nitrogen heterocycles (this figure rises to 16 if

cycles containing a benzene ring e.g. ergoline or tetrahydroiosquinoline are included).[58] A repeat of this analysis by Marshall *et al.* for drugs approved from January 2013 to December 2023 (after the publication of the Vitaku *et al.*'s analysis) found that 82% of these contained nitrogen heterocycles, with pyridine the most frequent (appearing in 54 of the 321 unique drugs, displacing piperidine in the previous analysis).[79] In the more recent analysis, 18 of the top 25 nitrogen heterocycles were aromatic (rising to 19 when tetrahydroisoquinoline is included), suggesting that the inclusion of aromatic nitrogen heterocycles in drugs is becoming more popular. A similar analysis by Delost *et al.* of oxygen-containing heterocycles in drugs approved up to December 2017 featured six aromatic heterocycles in the top 25 (the inclusion of benzene-containing partially saturated cycles lifts this figure to eight), although the inclusion of the sugars pyranose and furanose account for over 30% of all oxygen heterocycles, and thus skew these results somewhat.[80] Figure 1.7 demonstrates a selection of these top-represented aromatic heterocycles.

It is likely that the popularity of these aromatic heterocycles in medicinal chemistry campaigns is driven by a combination of the structural requirements of the target and the synthetic ease with which heterocycles and their derivatives can be accessed. As is discussed in the subsequent sections, aromatic heterocycles are well-suited to engage key molecular interactions with target binding sites and are thus frequently utilised to improve potency and selectivity in early stage campaigns.[81] However, the large corpus of synthetic methodologies for their synthesis and functionalisation has lead to aromatic heterocycles being well represented in screening libraries, which will have also contributed to their frequent inclusion in drug molecules.[82,83] These observed patterns thus likely reflect a combination of biological context-driven design and synthesis-driven bias in library construction.

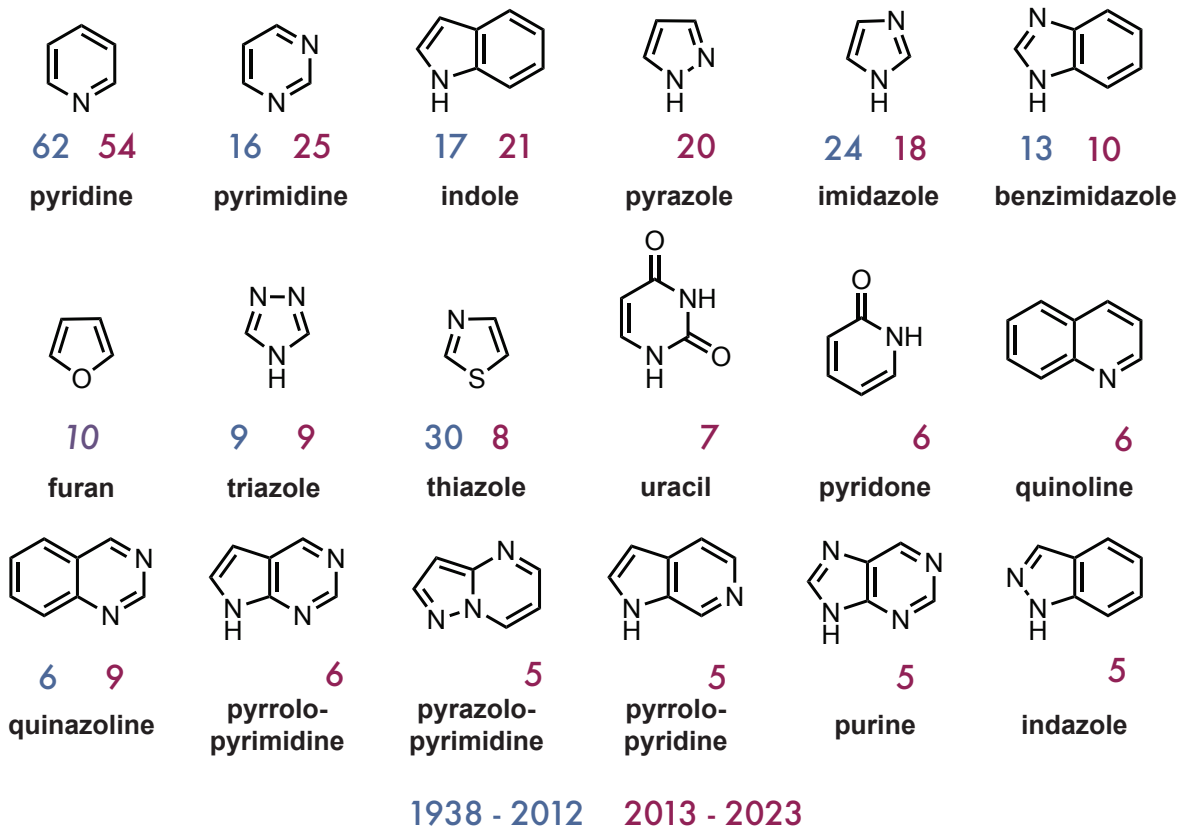


Figure 1.7. A selection of the top aromatic heterocycles in FDA-approved drugs, ordered by frequency of appearance in the 2013-2024 rankings. The number in blue indicates the number of unique drugs each heterocycle appears in from 2013-2024, and the number in red from 1938-2012. Furan's figures were taken from a separate study from 1938-2017.[58,79,80]

Inspection of Figure 1.7 reveals that a wide variety of aromatic heterocyclic architectures are regularly included in drug molecules. Although nitrogen overwhelmingly dominates as a heteroatom, oxygen- and sulfur-containing heterocycles are well represented, and there is a clear range of five-membered, six-membered, and fused ring systems with a wide-ranging heteroatomic constitution.

There are several properties of aromatic heterocycles that render them favourable to inclusion in bioactive molecules, including the ease with which H-bond donating and accepting atoms can be positioned at fixed geometries about a molecule, the flexibility with which the physic-

ochemical properties of the parent molecule can be adjusted by altering the constitution of a heterocycle, and the metabolic stability that aromatic heterocycles can afford.[81,84,85]

1.2.1 Hydrogen Bonding

Hydrogen bonds are a crucial instrument in the portfolio of non-covalent intermolecular interactions that govern the binding of small molecules to biological targets, and are recognised as the most important in determining a compound's selectivity and specificity.[86,87] It is recognised that the strength of H-bonding interactions are highly geometrically dependent, with studies determining that the strongest H-bonds form between a donor and acceptor with a distance in the range $2.7 - 3.0$ Å and at an angle close to $180°$.[86,88] Their rigid geometry, and the relative ease with which H-bond donors and acceptors can be selectively positioned around a ring to form H-bonds with target residues at a favourable orientation, have lead to aromatic heterocycles being widely deployed within small molecules for improving target specificity and selectivity.[86,87,89]

An illustrative example of the utility of aromatic heterocycles in selectively forming H-bond interactions with target proteins is the use of hinge-binding moieties in the design of kinase inhibitors.[89] Kinases are popular targets in oncology drug discovery, and since the approval in 2001 of **imatinib** (see Section 1.1), over 76 kinase inhibitors have since received FDA-approval for oncology and other indications.[38] Inhibitors have been categorised into four types based on their binding mode, of which the type I inhibitors make up over 90% of the approved drugs and, like the type II inhibitors, compete for association with the ATP-binding pocket.[89,90] These type I inhibitors all form H-bonds to residues in the hinge region, a flexible backbone connecting the N-terminal and C-terminal lobes of the kinase domain

which forms the ATP-binding pocket.[91] H-bonding interactions with the kinase hinge is essential for ATP-competitive ligands to bind potently.[92]

The natural ligand uses adenine to form H-bonds to the hinge backbone (illustrated in Figure 1.8 A for ATP binding to cyclin-dependant kinase 2), and other aromatic heterocycles have been widely deployed to mimic this interaction in small molecule inhibitors.[91,93] An analysis of the Pfizer crystallographic database by Xing *et al.* in 2015 identified 598 unique hinge-binding scaffolds from nearly 4000 kinase inhibitors in the database.[91] Although the authors do not quantify the prevalence of heterocycles within their results, every scaffold in the top 10 most frequent contains an aromatic heterocycle. A selection of these top 10 scaffolds are illustrated in Figure 1.8 B. The authors note that although the adenine moiety (which is found in the endogenous ATP ligand) is most commonly included in small molecule inhibitors, other heterocyclic cores are frequently included in its place to account for its lack of exit-vectors for further derivatisation, or to improve on selectivity.[91] In these applications, aromatic heterocycles are able to project H-bonding atoms (both donors and acceptors) in an appropriate orientation for hinge-binding, whilst also offering vectors for chemically bonding further functionality in a geometrically rigid manner.

Another illustrative example of the utility of aromatic heterocycles in selective H-bonding interactions is in the development of selective inhibitors of cathepsins S and L (Figure 1.9), cysteine proteases implicated in arthritis and cancer.[94] A screening campaign at AstraZeneca in the UK identified **1.1** as a dual-inhibitor of cathepsins S and L with sub-micromolar potency for both.[95,96] Analysis of the crystal structure of **1.1** bound to cathepsin L revealed that the backbone N-H bonds of Met70 and Asp71 were proximal to each other and aligned in parallel (see Figure 1.9 panel B). Wishing to engage H-bonds with both of these residues, a series of analogues of **1.1** with heterocycles bearing two adjacent H-bond acceptors appended

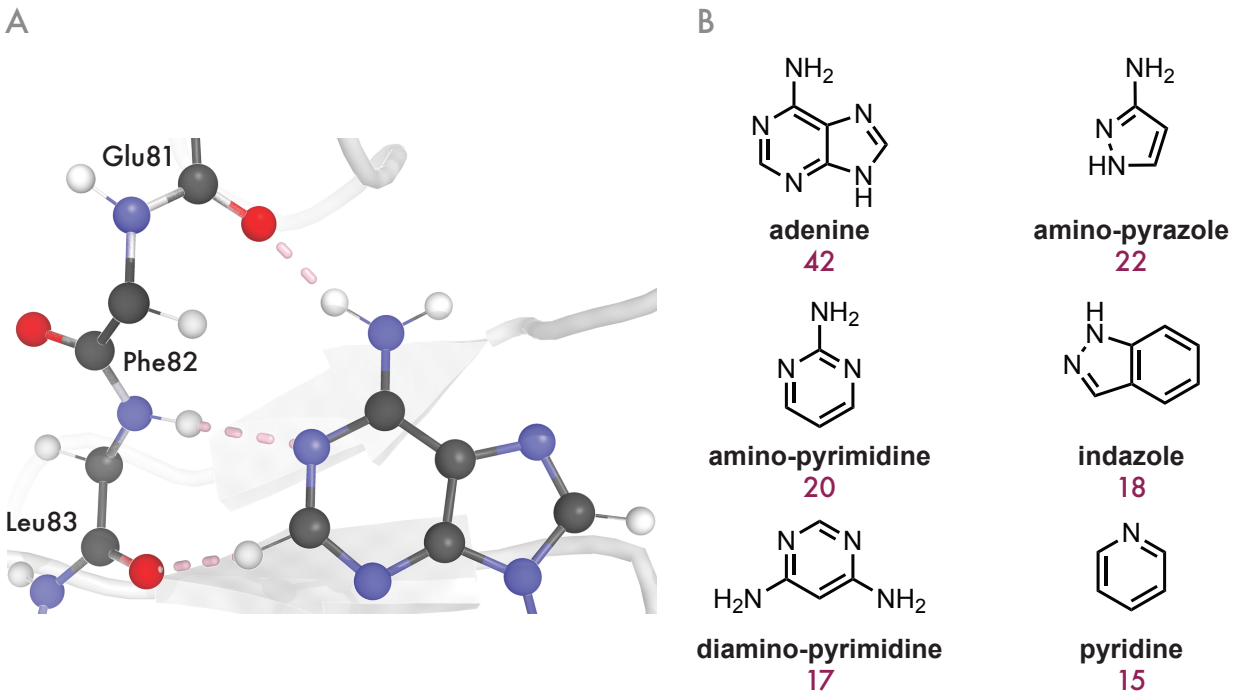


Figure 1.8. **A** | The hinge-binding interactions of adenine with the backbone peptidic chain in CDK2 (PDB: 8FP5).[93] **B** | A selection of the most frequent hinge-binding scaffolds from 598 unique scaffolds extracted from nearly 4000 kinase inhibitors in the Pfizer crystallographic database by Xing *et al.*[91] The number of times each scaffold is present in the hinge-binding region of a kinase inhibitor is also shown.

at the C3 position were synthesised and assayed (including 1,3,4-thiadiazole and pyridazines), with **1.2** being the most potent against cathepsin S with single digit potency, which was attributed to the ability of **1.2** to form dual H-bonds in close proximity with the binding pocket. None of the assayed heterocycles had a significant effect on cathepsin L potency, which the authors attributed to the fact that Leu69 in cathepsin L restricts the ability of the heterocycle to engage in H-bonding with the target residues more than the equivalent Phe69 in cathepsin S.[96]

A further benefit to the inclusion of an aromatic heterocycle at the C3 position is the synthetic ease with which these can be reliably further functionalised.[81] The authors also wished to introduce a salt-bridge with the charged side chain of Asp71, and the introduction

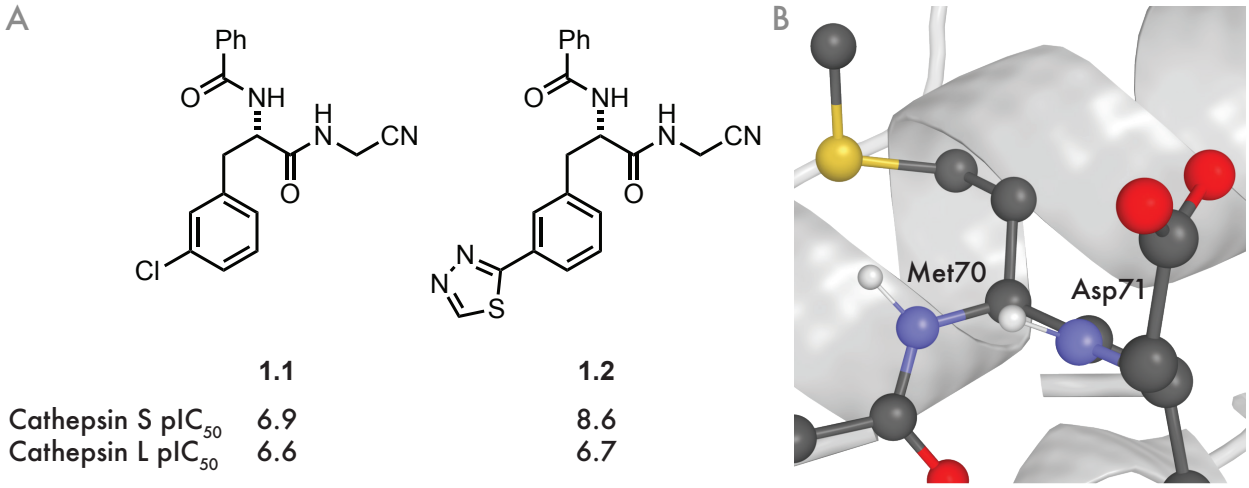


Figure 1.9. **A** | Two dual inhibitors of cathepsins L and S discovered at AstraZeneca.[95,96] **B** | The key residues in the binding pocket of cathepsin L (PDB: 3HHA). Key hydrogen atoms have been added for illustrative purposes.

of a methylamino moeity at C5 of the thiadiazole ring was projected in the correct orientation to engage this interaction in cathepsin L.[95]

1.2.2 Physicochemical Property Optimisation

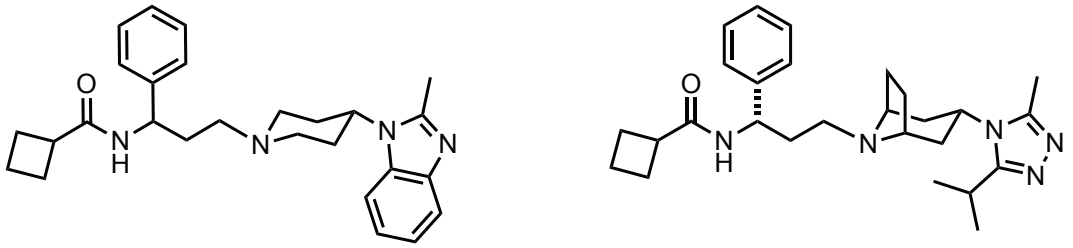
A favourable set of physicochemical properties is crucial for the success of a drug.[97] As previously mentioned, optimal ADMET properties are highly correlated with success in clinical trials, and the physicochemical properties of a candidate compound are crucial in determining its ADMET profile, thereby influencing its efficacy and safety.[64,97] Some key physicochemical properties that influence the pharmaceutical utility of a compound are the molecular weight, lipophilicity (*vide infra*), aqueous solubility (the amount of crystalline compound that remains in aqueous solution under equilibrium conditions), the number of H-bond donors and acceptors (as discussed above), and the molecular charge at physiological pH 7.4.[97–99]

Lipophilicity is a representation of a molecule's preference for existing in non-polar environments (e.g. fats or oils) over polar ones (e.g. water), and is commonly represented by the logarithm of the partition coefficient log P which measures experimentally a molecule's tendency to partition between *n*-octanol and water.[100] Whilst log P is an experimentally determined measure of lipophilicity for a neutral compound, its value can be influenced by ionisable groups within a molecule.[101] To account for this, the distribution coefficient log D is often measured, which reflects the compound's effective partitioning at a given pH (usually pH = 7.4) thus taking into account the presence of ionised and neutral species. In silico predictions of these values are commonly used in drug discovery, with cLogP and cLogD referring to the computationally predicted values of log P and log D, respectively.[102]

Achieving the right balance between aqueous solubility and lipophilicity is often viewed as a central factor in the success of a drug molecule.[103] Aqueous solubility determines the absorption of an orally-delivered drug into the bloodstream, and thus governs the exposure to the target; a poorly-soluble molecule will not be absorbed into the bloodstream through the gut and thus will not reach its intended target at therapeutically relevant concentrations, but a compound with too high an aqueous solubility can have a poor permeability, and thus will not cross the lipid bilayer of the gut membrane.[99,104] A degree of lipophilicity is necessary for a successful drug molecule to enable it to passively diffuse across lipid-based cell membranes, and to bind to protein targets which are often hydrophobic in nature.[102] However, too high a lipophilicity is linked to poor ADMET properties including high clearance, high toxicity, and poor *in vivo* selectivity.[69,102,105] It has therefore been proposed that an understanding of the factors influencing the lipophilicity of a lead molecule is the key descriptor for optimising a compound.[106]

As lipid environments are hydrophobic, increasing the hydrophobicity of a molecule will increase its lipophilicity, and increasing the polarity or adding ionic interactions will lower its lipid solubility. Both the polarity and the hydrophobicity arise from anisotropies in the electron distribution (how unevenly electrons are distributed about the molecule), therefore tuning these parameters involves altering the electronic distribution. Aromatic heterocycles therefore present an excellent opportunity for selectively modulating lipophilicity and solubility, as the asymmetric positioning of heteroatoms about a ring system can be used to control precisely the level of anisotropy in the electronic distribution, thereby influencing the lipophilicity of the parent molecule.[107]

These useful properties have lead to aromatic heterocycles being deployed to modulate lipophilicity and solubility in many medicinal chemistry projects.[81,97,108–112] An interesting example of this is in the development of the CCR5 chemokine receptor antagonist maraviroc as an anti-HIV therapy by Pfizer.[108] Although compound **1.3** displayed potent activity in inhibiting replication of HIV replication, with an IC_{50} of 75 nM, the need for life-long medication in HIV infection and the desire to maintain a plasma concentration of the drug well above the antiviral IC_{90} at all times meant that any candidate molecule required a large therapeutic window. Concerns about toxicity arising from potential hERG inhibition (a potassium channel in cardiac tissue, inhibition of which is linked to life-threatening cardiac arrhythmia) and high clearance rates lead to a campaign to reduce the lipophilicity of **1.3** without significantly increasing its molecular weight.[113] A previous hit in the initial screening campaign contained a 1,2,4-triazole moiety, and thus this higher-polarity heterocycle was incorporated in place of the methyl benzimidazole to form **1.4**, which displayed increased potency and significantly reduced lipophilicity when compared to **1.3**.



	1.3	**1.4**
Anti-viral IC_{50}:	75 nM	8 nM
cLogP:	3.2	1.6

Figure 1.10. The substitution of benzimidazole for 1,2,4-triazine reduces the lipophilicity of the parent molecule in the development of maraviroc.[108]

Another illustrative example of the utility of heterocycles in modulating solubility is in the use of isomers of oxadiazole.[114,115] It is found that 1,3,4-oxadiazole has significantly lower lipophilicity, higher aqueous solubility, and better metabolic and safety profiles than its 1,2,4-isomer, which is explained by a significant difference in dipole moment caused by greater anisotropy in its electron distribution.[115] This was utilised in a campaign to develop weight-loss therapeutics by scientists at AstraZeneca, who sought to identify small molecule antagonists of the Melanin Concentrating Hormone Receptor 1 (MCHr1), a G-protein coupled receptor (GPCR) predominantly found in the central nervous system (CNS; see Figure 1.11).[112,116] As MCHr1 resides in the CNS, any successful small molecule would need to reach the CNS from the bloodstream to engage the target receptor. The CNS is protected from external toxins by the blood-brain barrier, the existence of which restricts the physicochemical property space available to small molecules wishing to cross it, in particular with LogP tending to sit in the interval $[1, 3]$ and molecular weight < 450 Da.[117]

3-phenyl-1,2,4-oxadiazole **1.5** was found to have a lipophilicity of 3.1 (outside of the desirable range) and a very low solubility. Wishing to reduce the lipophilicity to within the desirable range for CNS penetration and improve the solubility, the 1,2,4-oxadiazole was subsituted for

its 1,3,4-isomer **1.6**, resulting in improved potency, a unit-improvement in lipophilicity and a 10 000-fold improvement in solubility, solely by the isomeric rearrangement of a nitrogen atom in a ring.[112]

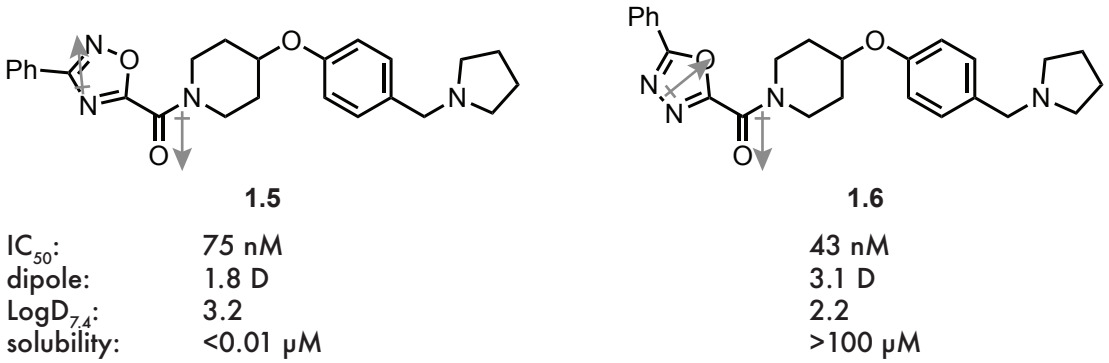


Figure 1.11. The substitution of a 1,2,4-oxadiazole for a 1,3,4-oxadiazole to reduce lipophilicity and improve solubility as part of an AstraZeneca campaign to improve the CNS exposure of MCHr1 antagonists for weight-loss.[112] The dipole values shown are for the 5-membered aromatic heterocycles, and some of the key dipole interactions are illustrated.

The sensitivity of the physicochemical properties of aromatic heterocycle-containing compounds to the number and position of heteroatoms with the ring system was illustrated nicely by Cosgrove *et al.*[118] To explore the effect of rearranging nitrogen atoms within the aromatic core scaffold of a typical candidate compound in medicinal chemistry, a matched molecular series (MMS) of compounds was prepared where the R-groups are held constant and the cores differ only in the position or number of aromatic nitrogen atoms (Figure 1.12).

In keeping with the pattern from the previous examples, the key physicochemical properties depend strongly on the constitution of the heterocycle involved. Although it is difficult to directly discern patterns, it is clear that **1.8** with the highest lipophilicity has the lowest solubility, but compounds **1.9** – **1.11** all have similar lipophilicity yet varying degrees of solubility. Of note is that the lipophilicity of **1.8** can be reduced by a unit in **1.11** (which corresponds to a power of ten difference in partition coefficient) with an increase in molecular

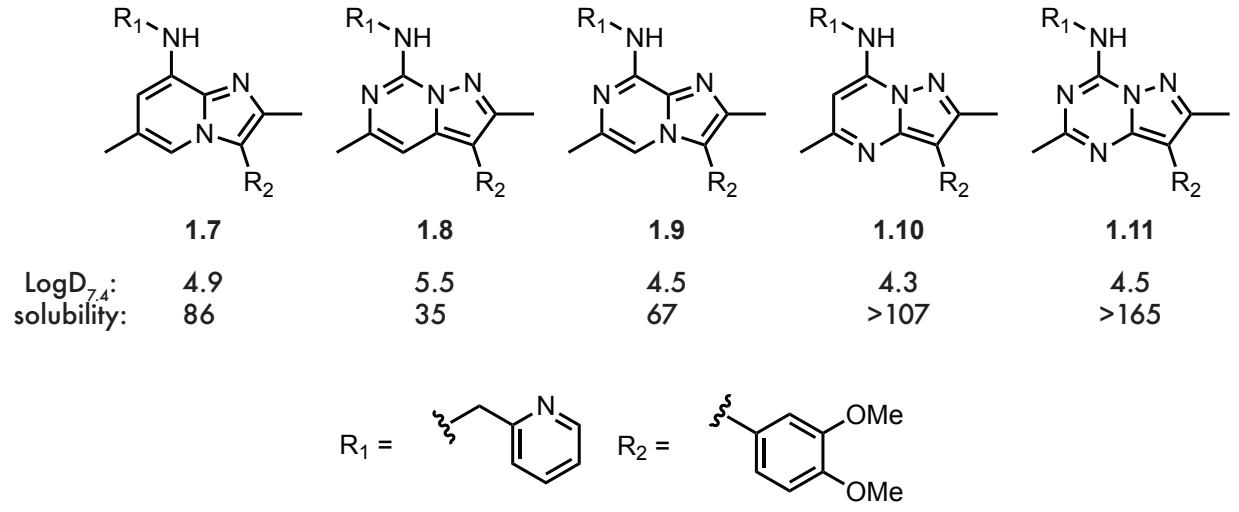


Figure 1.12. The effect of constitutional isomerism of 5,6-bicyclic heterocycles on kinetic solubility in a typical lead-like compound.[118] Solubility is measured as a kinetic solubility at pH = 7.4 in $\mu g\,mL^{-1}$.

weight of just one Dalton. This illustrates the flexibility of aromatic heterocycles in drug design, where subtle structural changes can profoundly impact physicochemical properties, enabling fine-tuning of drug-like characteristics with minimal structural perturbation.

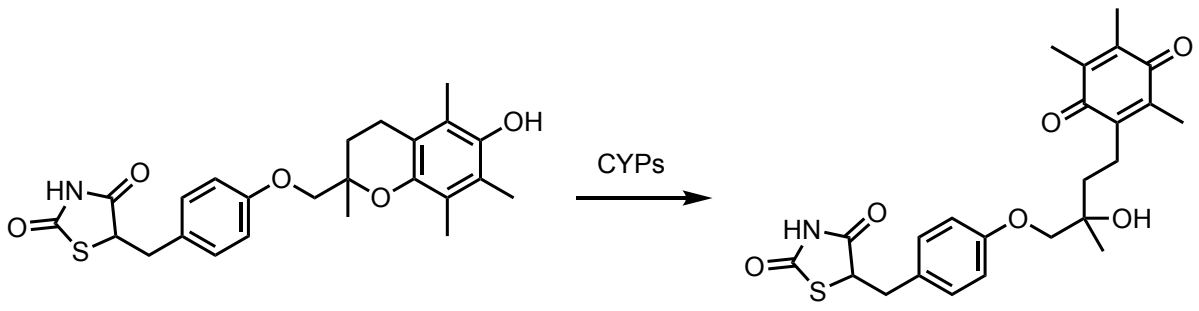
1.2.3 Improving Metabolic Stability

Metabolism is the process by which biotransformation enzymes (most prevalent in the liver) increase the polarity of foreign molecules to promote their excretion through the urinary or biliary system.[119,120] It is classically separated into two chronologically distinct chemical processes: phase I and phase II metabolism. Phase I metabolism involves the oxidative functionalisation of compounds to increase their polarity, and is primarily carried out by haem-containing cytochrome P450 enzymes in the liver, of which the most prevalent CYP3A4 is responsible for the metabolism of around 45% of all known drugs.[119,121,122] Typical phase I reactions include *N*- and *0*-dealkylation, aliphatic and aromatic hydroxylation, *N*- and *S*-oxidation, and deamination.[123] Phase II metabolism transforms compounds, often

after phase I metabolism has increased their polarity, to more easily excretable forms by conjugating polar groups to hydrophilic moieties including glucuronic acid, glutathione, and sulfonyl groups, thereby further increasing the polarity of the molecule and encouraging its elimination.[119,123–129] Metabolism is the principal factor governing the deactivation, toxification, and detoxification of small molecules and thus is key in determining their half-life and safety profile.[130] Furthermore, although the evolutionary role of metabolism is to remove harmful xenobiotics from organisms, the metabolites generated in phase I and II processes can also be chemically reactive, and toxic metabolites can trigger serious adverse drug reactions leading to failure at clinical trials and occasional withdrawal of previously-approved drugs.[131,132]

An example of the danger of toxic metabolites is the antidiabetic and anti-inflammatory drug **troglitazone** (Figure 1.13), which received FDA-approval in January 1997.[133] However, evidence of liver failure and other adverse hepatic events in patients taking **troglitazone** eventually lead to the FDA revoking its approval in March 2000, with the subsequent market withdrawal estimated to have cost Glaxo Wellcome in excess of $130 million USD.[133,134] Although the precise causes of the observed hepatotoxicity is uncertain, the quinone metabolite **1.12** was identified and attributed to oxidation by the phase I liver enzymes CYP3A4 and CYP2C8.[133,135] Quinones are known to be toxic metabolites of many drugs containing phenol-like functionality, and it was thus proposed that this product of metabolism contributed to the toxicity leading to the drug's withdrawal.[133,136]

It is thus of significant importance that the metabolic profile of lead-like compounds is carefully monitored and adjusted if a drug discovery project is to be successful.[132,138] The susceptibility of a compound to metabolism by CYP enzymes is largely determined by the set of physicochemical properties discussed in Section 1.2.2, and thus strategies to reduce



troglitazone **1.12**

Figure 1.13. Metabolism of the antidiabetic drug troglitazone by liver enzymes CYP3A4 and CYP2C8 generated the proposed hepatotoxic quinone metabolite **1.12**.[137]

the metabolic liability of compounds usually involve reducing the overall lipophilicity of the molecule.[119,139]

A common experimental approach for assessing metabolic stability involves incubating compounds with vesicular fractions derived from postmortem livers that contain key drug-metabolising enzymes (particularly cytochrome P450s) called microsomes, of which human liver microsomes (HLMs) and mouse liver microsomes (MLMs) are the most common. The rate at which a compound is metabolised within these microsomes provides an estimate of its intrinsic clearance (Cl_{int}; a measure of the liver's ability to eliminate a drug in the absence of blood flow limitations and protein binding) and can be used to calculate its *in vitro* half-life ($t_{\frac{1}{2}}$; the time taken for the concentration of the substance to reduce to half its initial value). Compounds with longer half-lives in microsomal assays are typically considered more metabolically stable, and such data are frequently used to guide structure modification during lead optimisation.

As the techniques for metabolic profiling of compounds became more sensitive, including the use of high resolution mass spectrometry and NMR spectroscopy, it became possible to identify a molecule's metabolic 'soft-spots': labile sites where phase I oxidation reactions take

place.[120,140,141] These sites often correspond to regions of the molecule where hydrogen abstraction is sterically facile, areas of higher electron density (for example allylic and benzylic positions), and heteroalkly groups, which frequently undergo dealkylation reactions.[84,142] An example of the metabolic 'soft-spots' of the HIV protease inhibitor **indinavir** is shown in Figure 1.14.[132,143] The extensive metabolism of **indinavir** results in its short half-life of just two hours, and thus the requirement that it be dosed every eight hours.[132]

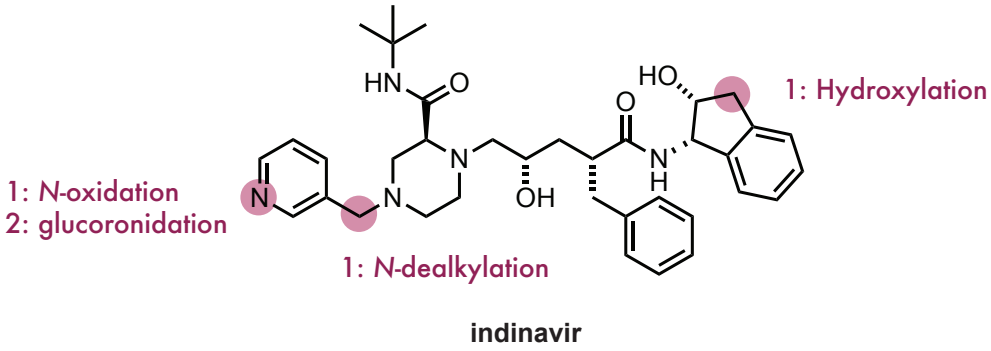
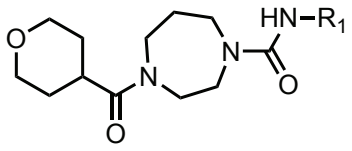


Figure 1.14. The metabolic soft-spots of the HIV protease inhibitor indinavir. The numbers indicate whether the site is prone to phase I or phase II metabolism.[132,143]

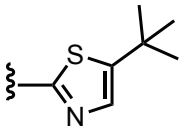
The same properties that make aromatic heterocycles so useful in the modulation of physicochemical properties and the selective formation of strong non-covalent intermolecular attractions discussed previously also describe their utility in modulating the metabolic stability of parent molecules. Metabolic 'soft-spots' on aromatic systems can often be attenuated by altering the electronic properties of the ring to reduce the energy of the HOMO and alter the electronic distribution; a task which often is as theoretically straightforward as rearranging the ring heteroatoms.[84]

In a campaign to develop agonists of the GPCR cannabinoid receptor CB2 for anti-inflammatory indications, a high-throughput screening (HTS) at Boehringer Ingelheim identified a 1,4-diazapane scaffold, which was further optimised for selectivity over the CB1 receptor and HLM stability to give the scaffold **1.13** (Figure 1.15).[144,145] *Tert*-butyl thiazole **1.13a** was

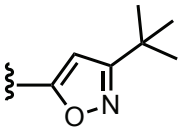
subsequently identified, which had a lipophilicity and potency within the desired range, but a moderate stability (HLM $t_{\frac{1}{2}} = 27$ min). By exchanging the thiazole ring for isoxazoles **1.13b** and **1.13c**, the stability was significantly increased (HLM $t_{\frac{1}{2}} > 120$ min).[145] Interestingly this improvement in metabolic stability cannot be explained by lipophilicity, as **1.13b** and **1.13c** both have higher $cLogD_{7.4}$ values than **1.13a**, and **1.13c** is significantly more lipophilic than **1.13b** despite showing similarly high microsomal stability. It was postulated that the isoxazoles are less electron-rich than the thiazole, as indicated by the more negative ionisation potential, and as such are less susceptible to oxidative phase I metabolic reactions than the thiazole equivalents.[144] It is also interesting to note, in light of the discussion in Section 1.2.2, that switching the positions of the oxygen and nitrogen atoms in the isoxazole ring leads to significant differences in lipophilicity between **1.13b** and **1.13c**.



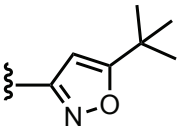
1.13



a



b



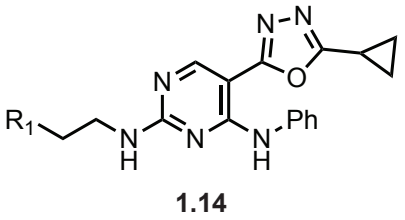
c

	a	b	c
EC_{50} (nM)	150	26	67
HLM $t_{1/2}$ (min)	27	>120	>120
$cLogD_{7.4}$	0.77	0.98	1.75
I.P. (eV)	9.5	10.2	10.2

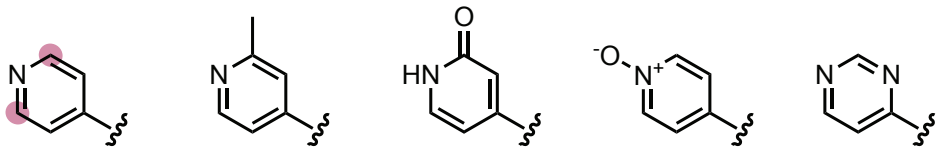
Figure 1.15. Riether *et al.*'s work to increase the metabolic stability of agonists of the CB2 receptor by switching a thiazole ring (**1.13a**) for an isoxazole (**1.13b-c**).[144,145]

Another illustrative example of the utility of aromatic heterocycles in tuning metabolic stability comes from Ishida *et al.*'s work to discover small molecule inhibitors of FMS-like receptor tyrosine kinase 3 (FLT3) for the treatment of acute myeloid leukaemia.[146] HTS

investigations identified a 5-(1,2,4-oxadiazol-2-yl)pyrimidine derivative as a primary hit, and further structure-activity relation (SAR) work identified pyridine analogue **1.14a** as a potent FLT3 inhibitor, with a growth inhibition 50% (GI_{50}) value in an MOLM-13 tumour cell line model of 40 nM. However, metabolic profiling of **1.14a** showed it to be unstable in both human and mouse liver microsomes, with a high Cl_{int} of 19 $L\,h^{-1}\,kg^{-1}$. Metabolite analysis revealed that the 2-position of the pyridine (highlighted in red in Figure 1.16) was the primary site of oxidative metabolism, but attempts to block this site with a methyl group without altering its lipophilicity (**1.14b**) had no effect on the metabolic stability. When the polarity of the ring was increased by addition of pyridone **1.14c**, pyridine *N*-oxide **1.14d**, or pyrimidine **1.14e** the metabolic stability was greatly increased.[144,146] Although **1.14d** and **1.14e** resulted in an increase in GI_{50}, pyridone **1.14c** maintained sufficient potency at significantly increased metabolic stability, and with a low lipophilicity.



1.14

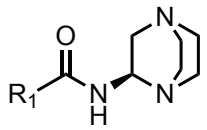


	a	b	c	d	e
GI_{50} (nM)	40	137	113	3610	775
HLM Cl_{int} (L/h/kg)	19	22	2.5	0.2	4.4
$cLogD_{7.4}$	1.49	1.92	0.32	-0.42	0.44

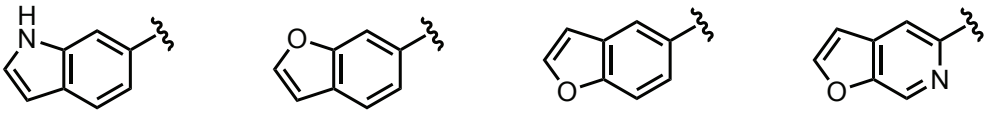
Figure 1.16. Improving the metabolic stability of FLT3 inhibitors by increasing the polarity of the aromatic heterocyclic ring.[144,146] The atoms highlighted in red indicate the suspected site of metabolic oxidation.

Fused aromatic heterocycles can also be adjusted to improve metabolic stability. Seeking to develop small molecule therapeutics for schizophrenia, a group at Pfizer performed a high-

throughput screen for agonists of the α7 neuronal nicotinic acetylcholine receptor (nAChR), identifying quinuclidine amide indole **1.15a** as a potent agonist with an $EC_{50} = 100$ nM.[144,147] SAR about the aromatic heterocyclic portion identified benzofurans **1.15b** and **1.15c** with similar potencies and improved microsomal stabilities (Figure 1.17). The investigators were concerned about the tendency for benzofurans to undergo metabolic activation *in vivo*, and thus sought to reduce electron density at the benzene ring with furopyridine analogue **1.15d**, which had a similar potency, a slightly reduced microsomal stability, and a significantly reduced lipophilicity.[148] Reactive metabolite assays on **1.15b–c** showed **1.15b** to be positive, although interestingly regioisomeric **1.15c** was negative.



1.15



	a	b	c	d
EC_{50} (nM)	100	86	47	65
RLM[a] (%)	14	69	77	55
RMA	N.D.	Positive	Negative	Negative.
$cLogD_{7.4}$	0.01	0.52	0.52	-0.52

Figure 1.17. Wishka *et al.*'s improvement of the metabolic stability of α7 nAChR agonists by modulating the structure of the aromatic heterocycle.[144,147] [a] Metabolic stability is measured *in vitro* in rat liver microsomes, and is expressed as the percentage of the compound remaining after one hour. RMA: reactive metabolite assay.

1.2.4 Perhaps Flatland isn't so bad

In their now-infamous (at the time of writing it had been cited 3211 times) 2009 paper 'Escape from Flatland', Lovering *et al.* correlated the success of compounds through the

stages of clinical trials with the fraction of carbon atoms in the molecule that are sp^3 hybridised.[149] Arguing that an increase in the saturation of candidate molecules would allow for greater molecular complexity, and thus expand access to more favourable areas of chemical and physicochemical space, they pushed for a shift in the focus of drug discovery campaigns away from aromaticity and towards the inclusion of a greater proportion of three-dimensional molecular complexity in leads.

They found that the average Fsp^3 (the proportion of carbon atoms in the molecule that are sp^3 hybridised) was 0.36 for medicinal chemistry 'discovery' compounds, but that this figure rises to 0.47 for approved drugs, and that through each phase of clinical trials, the Fsp^3 increases at a statistically significant level.[149] There appeared to be a significant link between the chances of a molecule's clinical success, and its degree of saturation, despite the statistic revealing that an average successful, FDA-approved drug has more than half of its carbon atoms in sp^2 hybridisation.

Revisiting the analysis of Lovering *et al.* 15 years later, Churcher *et al.* compared the Fsp^3 of approved drugs and those in clinical trials both before and after 'Escape from Flatland' was published.[150] Their data, displayed in Figure 1.18, suggests that the conclusions reached in the original paper have not persisted since its publication, with the average Fsp^3 of drug molecules approved since 2009 (shown in blue in Figure 1.18 panel A) appearing to decrease compared to those approved before 2009. Extending this analysis to molecules in various phases of clinical trials in mid-2024, they found no clear relationship between the highest phase reached by a molecule and its Fsp^3 (Figure 1.18 panel B), also noting that there existed across all datasets multiple examples at the extrema, with Fsp^3 values of 0 and 1.[150] The authors are hesitant to suggest reasons for this trend since 2009, but postulate that it could be related to drug target selection trends, citing as an example the increase in the

approvals of kinase inhibitors (see Section 1.2.1), and also the rise in scope and popularity of metal-catalysed cross coupling reactions.[150,151]

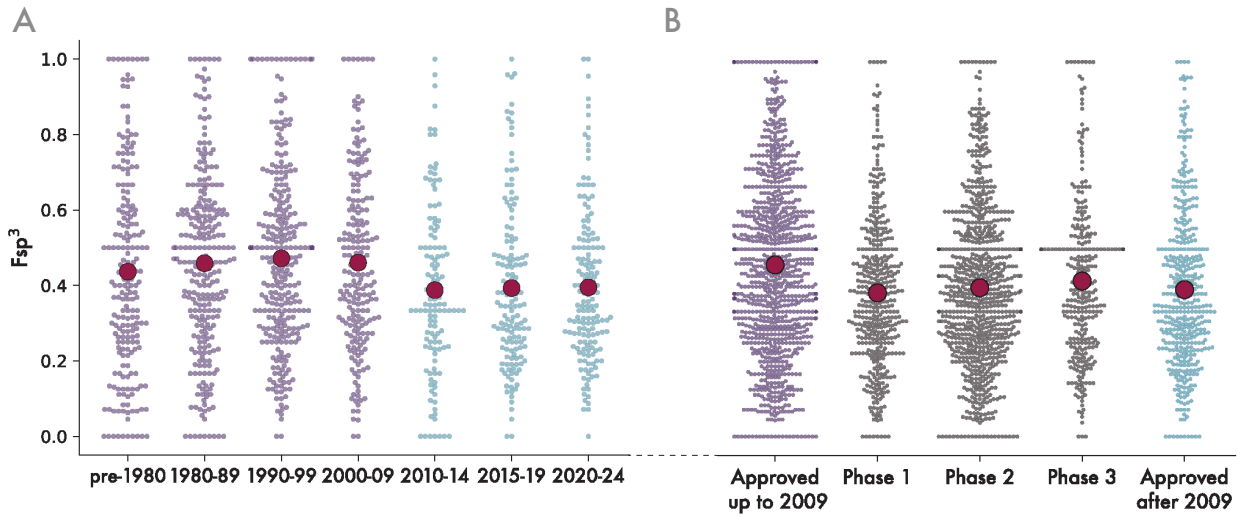


Figure 1.18. **A** | The distribution of Fsp^3 values for FDA-approved drugs, categorised by the year of approval. Drugs approved before Lovering *et al.* are shown in purple, and those after in blue. The mean Fsp^3 for each interval is represented by the red circle. **B** | The Fsp^3 distribution for drugs approved before 2009 (purple), after 2009 (blue), and those in the designated phases of clinical trials in mid-2024 (grey). Mean Fsp^3 values are shown as red circles. This figure was made using data from Churcher *et al.*[150]

Churcher *et al.* suggest that Fsp^3 might not be a useful metric for optimisation in drug discovery campaigns, suggesting instead that achieving a balanced profile of physicochemical properties, on-target potency, and ADMET characteristics directly, rather than relying on simplified surrogate metrics, is likely to be the best way to identify successful drug molecules.[150] Although Lovering's original insights are valuable in highlighting potential advantages of increased molecular complexity, the subsequent shift back toward lower Fsp^3 compounds suggests that medicinal chemistry optimisation is context-dependent. Indeed the simple dichotomy between 'flat' and 'three-dimensional' is probably too great an abstraction, and the results of Churcher *et al.* suggest that success in drug discovery is better achieved with a balanced integration of multiple factors including target selection, synthetic

accessibility, and carefully optimised physicochemical properties. As has been discussed in the preceding sections, aromatic heterocycles offer unique opportunities for modulating these key parameters and therefore Flatland, despite its suggested limitations, remains an important and productive area of chemical space within contemporary drug design.[150]

1.3 Bioisosterism

1.3.1 Bioisosterism: Form Follows Function

There have existed numerous formal and informal definitions of bioisosterism, but it can conceptually be considered as the replacement of functionality in a molecule (be it a core scaffold or pendant functional groups) to generate analogues with similar pharmacodynamic properties, but improved physicochemical, pharmacokinetic, or synthetic properties.[152–156] It builds on the concept of isosterism originally formulated by Moir and Langmuir in the early parts of the 20th century, when the physical properties of chemically distinct compounds with the same number of electrons (for example N_2O and CO_2) were found to be remarkably similar.[97,154,157,158] Langmuir used this principle to tabulate a list of isosteres grouped by their total number of electrons, and used this to make predictions about unmeasured physical properties of some compounds (for example the freezing point of cyanic acid); some examples of this are shown in Table 1.1.[157] Grimm extended this concept with his hydride displacement law, which hypothesised that adding a hydrogen to an atom will result in a 'pseudoatom' with properties similar to those of an atom of the element with the next highest atomic number (therefore CH and N, or NH and O would share similar properties).[152,159,160]

This early notion of similarities in electronic structure being linked to observable physical properties was then extended to biological systems by Hans Erlenmeyer in a 1935 paper,

Table 1.1. A selection of Langmuir's isosteres.[157]

Type	Isosteres
1	H^-, He, Li^+
8	N_2, CO, CN^-
10	CO_2, N_2O, N_3^-, CNO^-
15	ClO_3^-, SO_3^{2-}, PO_3^{3-}

where he showed that antibodies were not able to discriminate between phenyl and thiophene rings, or molecules differing by O, NH, or CH_2 in artificial antigens created by derivatising proteins with diazonium ions, and was the first demonstration that structurally distinct molecules could have similar biological properties.[97,159] Since then, more contemporary definitions by Friedman in 1950 (*atoms or molecules which have the same type of biological activity*) and Burger in 1991 (see page 60) relate structural and electronic similarity to deliberately non-specific *biological effects*.[97,153,158] By relating structural similarity to biological effect, rather than directly to physicochemical properties as had been done previously, both Friedman and Burger's definitions recognise that bioisosterism is inherently context dependent; the success of a bioisosteric replacement depends on the desired properties seeking to be optimised, and the biochemical and physiological setting.[97] However, a judicious design and selection of bioisosteres can address significant issues in drug discovery campaigns, and thus bioisosterism has become a key strategy for developing drug candidates in medicinal chemistry.[97,152,155,161–165]

Arguably the simplest example of a bioisosteric pairing is the exchange of hydrogen atoms for deuterium. When forming intermolecular interactions with a target, the two atoms are virtually indistinguishable, and so their exchange should have little effect on the binding affinity at the target. However, deuterium has an atomic mass effectively double that of hydrogen, and thus the energy required to break the X–D bond is larger than that required for the X–H

bond, reducing the rate of any chemical reactions that involve breaking this bond (termed a kinetic isotope effect).[166] This becomes relevant when considering the metabolism of drug molecules, where hydrogen abstraction is often a crucial step in oxidative metabolism.[167] An example is given by Bristol-Myers Squibb's development of the oral allosteric TYK2 inhibitor **deucravactinib** for the treatment of psoriasis, which received FDA approval in 2022 (Figure 1.19).[168,169]

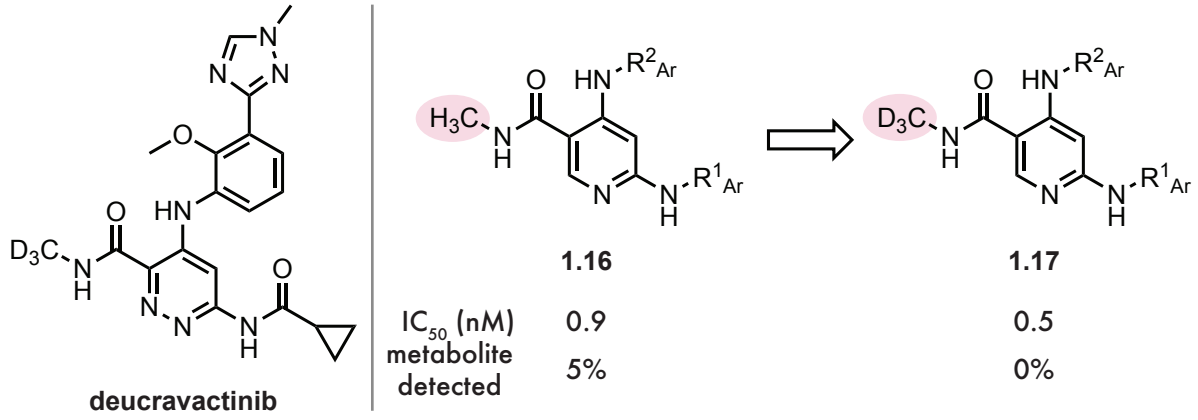


Figure 1.19. Hydrogen to deuterium bioisosterism in the development of **deucravactinib** for the treatment of moderate-to-severe plaque psoriasis. Deuteration of the methyl amide eliminated formation of the promiscuous primary amide metabolite, improving selectivity.[168–170]

Secondary amide **1.16** was identified as a potent inhibitor of TYK2 with a good selectivity profile, however metabolic studies revealed it was readily metabolised *in vivo* to the primary amide, which was a sub-micromolar inhibitor of 20% of tested kinases.[170] Deuteration at the methyl amide to give **1.17** reduced the rate of this metabolism to the extent that levels of non-specific primary amide metabolite were undetectable, and **1.17** maintained a highly favourable potency and selectivity profile.[168,170]

1.3.2 Bioisosterism in Medicinal Chemistry

Bioisosteres are often deployed at the lead-optimisation stage of a project, and frequently look to address issues in the ADMET profile of a lead whilst maintaining target engagement and potency.[152,171] Thornber proposed that the desired effect of a bioisosteric replacement strategy could generally be assigned to one or more of four categories:

1. **Structural**: projecting key target-interacting functionality at a fixed geometry. Bioisosteric replacement could improve the binding affinity by effectively pre-organising the ligand into an active conformation, or maintaining the key functionality at the required orientations.

   An interesting example of this is given in the development of potent inhibitors of Factor Xa for a haematology indication (Figure 1.20).[168,172] Previous work had identified cycloheptadienone **1.18** as a potent inhibitor that was liable to photoisomerisation to an inactive conformation. Bioisosteric substitution of the troublesome cycloheptadienone to give diphenoxypyridine analogue **1.19** rendered the ligand photochemically inert whilst maintaining potency, which was attributed to the strong preference for phenoxypyridines to exist in the *syn* rather than *anti* conformation to avoid a repulsive lone pair interaction (see Figure 1.20).[172]

2. **Target interaction**: forming important intermolecular interactions with the target binding site. Key parameters to consider when bioisosterically replacing functionality directly interacting with the target are the size and shape, electronic distribution including positioning of H-bonding atoms, and the pK_a.

3. **Pharmacokinetics**: improving the absorption, distribution, and excretion properties of the molecule. Lipophilicity, pK_a, and solubility are among the important parameters to consider.

4. **Metabolism**: increasing or reducing the metabolic liability of a compound.[h] Again lipophilicity is a key measure, but chemical and metabolic reactivity also need to be considered.[152,160,173]

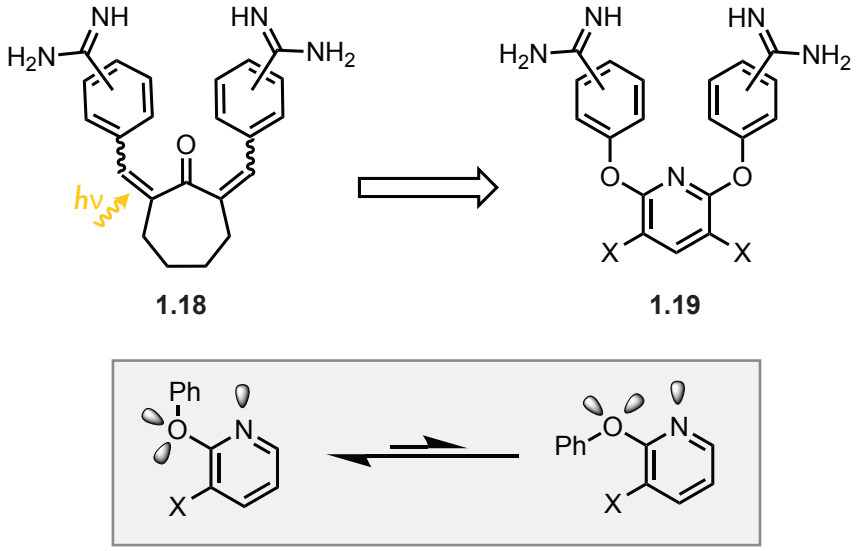


Figure 1.20. Bioisosterism to lock the conformation of an inhibitor of Factor Xa for the design of direct acting anti-coagulants.[174] Replacement of the cycloheptadienone core with a 2,6-diphenoxypyridine prevented photoisomerisation, and the strong preference for the *syn* conformation restricted the ligand to the active geometry.[168,172,174,175]

Although not one of Thornber's original categories, bioisosterism also finds use in taking projects into novel intellectual property space.[97,152,176] It is no secret that the development of successful drugs comes with significant financial reward, thus there are incentives for the pharmaceutical industry to develop 'me-too' or 'me-better' compounds in successful indications, for which the avoidance of legally-protected chemical space is an important

[h] Although more unusual, there are examples of using bioisosterism to introduce metabolic soft-spots to reduce half life; *vide* reference [173]

consideration.[177,178] Bioisosteric replacement therefore presents an opportunity for exploring novel, potentially patentable chemical space without significantly sacrificing progress made on an existing project.[179]

An example of several of the considerations mentioned above is in the development of the selective phosphodiesterase 5 (PDE5) inhibitor **vardenafil** by Bayer (Figure 1.21 panel A).[180] **Sildenafil** had previously been identified by scientists at Pfizer as a potent inhibitor of PDE5, whose indication had been changed from angina to treatment of male erectile dysfunction following reports of an unexpected side-effect in the phase 1 clinical trials.[181] However, sildenafil's limited selectivity over the structurally related PDE6 (expressed in the rod and cone photoreceptors of the eye) was known to cause temporary issues with colour vision in some patients.[181,182] Seeking to improve the potency and selectivity profile of **sildenafil**, scientists at Bayer performed a bioisosteric 'scaffold-hop' which improved the binding affinity by a factor of 10, whilst increasing the selectivity ratio for PDE5 over PDE6 to 16, and generating new intellectual property; factors which are illustrated in the recommended dose for **vardenafil** being one-fifth that of **sildenafil**.[179,183–188]

Another illustrative use of bioisosterism is in the replacement of carboxylic acids with tetrazoles. Carboxylic acids are important motifs in drugs as their low $pK_a \approx 4.2 - 4.4$ leads to them bearing a negative charge in physiological conditions, rendering them useful for forming salt-bridge interactions with positively charged residues in target binding pockets.[87] Such is their importance that in 2013 over 450 FDA-approved drugs were found to contain a carboxylic acid, with the number likely to have increased since then.[189] However, despite their prevalence in approved drugs, the presence of a carboxylic acid group has been linked to undesirable properties including reduced permeability, increased toxicity, and reduced metabolic stability.[191] The polar N-H bond of tetrazoles, coupled with extensive delocali-

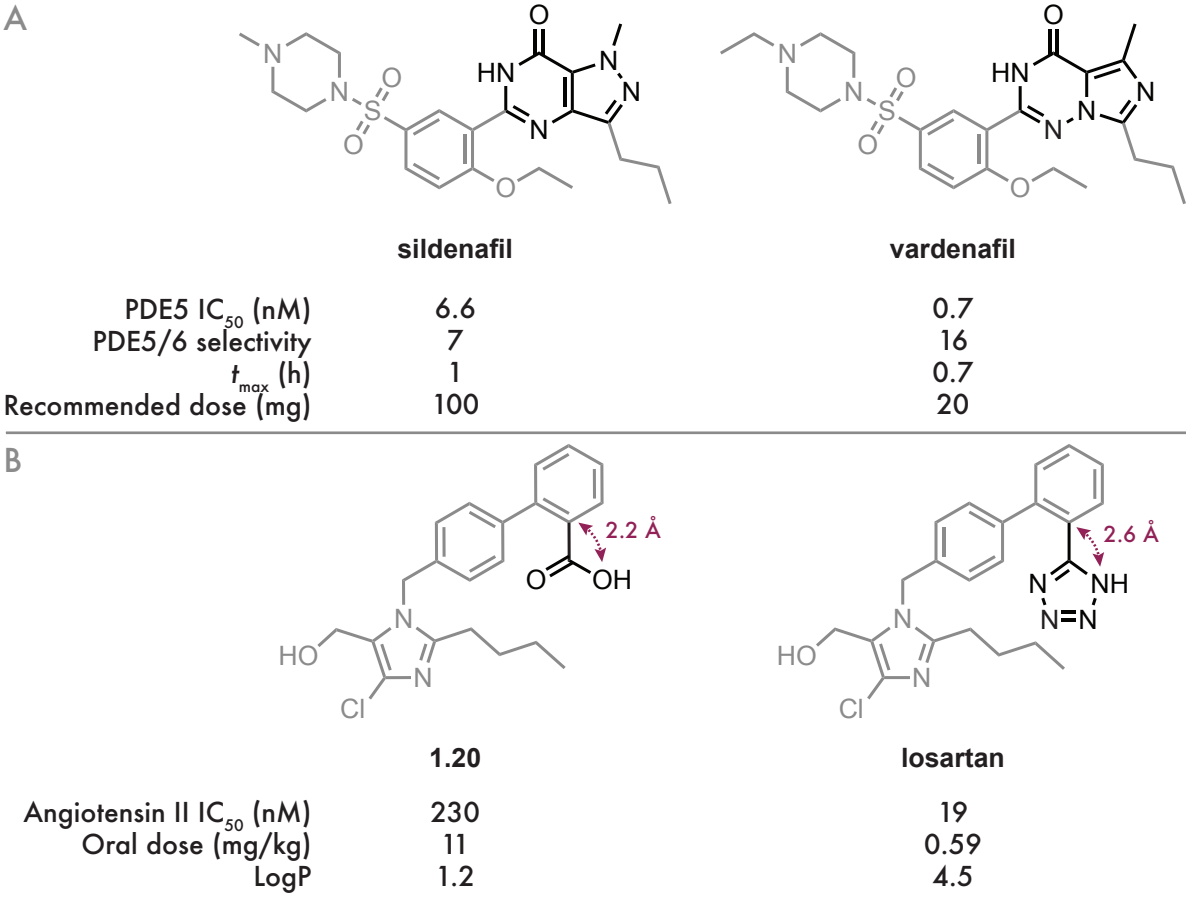


Figure 1.21. Examples of bioisosterism in FDA-approved drugs. **A** | A bioisosteric scaffold-hop of the central aromatic heterocyclic core of **sildenafil** gives **vardenafil** with improved pharmacological properties, resulting in a lower recommended dose.[184–188] **B** | Bioisosteric replacement of the pharmacologically troublesome carboxylic acid moiety in **1.16** with a tetrazole in the FDA-approved **losartan** improves the binding affinity and reduces the oral dose by a factor of ten.[189,190]

sation of the resulting negative charge into the aromatic system, makes them of a similar acidity to carboxylic acids ($pK_a \approx 4.5 - 4.9$) and both are planar in nature.[162] However, despite being of a similar acidity, anionic tetrazoles are almost 10 times more lipophilic than carboxylates, and have a lower charge density due to delocalisation into the aromatic ring.[162,192] These favourable properties, particularly the increase in lipophilicity with its corresponding improvements in cell permeability, have made tetrazoles popular bioisosteres of carboxylic acids in medicinal chemistry, with a search of the ChEMBL 28 database using

SwissBioisostere (*vide infra*) revealing that 521 individual carboxylic acid to tetrazole substitutions have been documented across 134 separate targets.[193,194]

An example of the application of this bioisosteric pairing is in DuPont's development of the non-peptidic selective angiotensin II receptor antagonist **losartan** (Figure 1.21 panel B).[195] **1.20** was initially identified as having promising *in vitro* activity against angiotensin II and was active in rats by intravenous injection, however its oral bioavailability was too low to progress the molecule as a drug candidate. Substitution of the C2 carboxylic acid for tetrazole gave **losartan**, which not only had a significantly improved oral bioavailability (contributing to a reduction in the required oral dose from 11 $mg\,kg^{-1}$ to 0.59 $mg\,kg^{-1}$), but also improved binding affinity at the receptor by a factor of 10.[162,189,195] The investigators attributed the improvement in oral bioavailability to the increase in lipophilicity (logP of 1.2 for **1.20** compared to 4.5 for **losartan**), and the increase in binding affinity to the distribution of negative charge about the ring, facilitating a stronger interaction with Lys199 and His256 in the binding pocket.[162,190]

1.3.3 Aromatic Heterocycles as Bioisosteres

The cases discussed above are but a few examples of what is a versatile and powerful technique for improving the overall pharmacological profile of a drug molecule.[97] It is outside the scope of this thesis to list examples of all the possible modes of bioisosteric replacement, however reviews and analyses by Burger 1991 (an excellent, comprehensive review of the early field), Patani 1996 (a broad overview), Meanwell 2011, Meanwell 2013, Meanwell 2018 (the use of fluorine in bioisosterism), Kumari 2020 (bioisosteres of amide bonds), Subbaiah 2021 (bioisosteres of phenyl rings), Meanwell 2023, Ertl 2023 (bioisosteres of the most com-

mon linking fragments) and Tsien 2024 (an overview of the emergence of sp^3 bioisosteres of benzene) cover the field thoroughly.[97,154,158,168,196–201]

The properties of aromatic heterocycles that favour their inclusion in small molecule drugs (as discussed in Section 1.2) also make them favourable as bioisosteres; specifically the ease with which small (often single atom) changes to their constitution can bring about significant alterations to their physicochemical properties, without disrupting interactions with their target. Indeed all of the examples in Section 1.2 could be considered examples of bioisosteric replacement. Given the pharmacological impact of such substitutions, there is a clear need for systematic methods to identify and prioritise potential bioisosteric pairings, especially as the number of possible replacements far exceeds what can be explored experimentally.[202]

1.4 Tools for Identifying Bioisosteric Pairings

Given the widespread use and advantages that deploying bioisosteres can confer on drug discovery campaigns, methods and tools for the systematic retrieval and identification of bioisosteric pairings have been developed both in the literature and behind-closed-doors in pharmaceutical firms. These can broadly be divided into two categories: those that identify pairings from existing biological databases or list previously-made pairings, and those that are able to predict novel pairings by searching through virtual compound libraries or using statistical approaches.

1.4.1 Pre-existing Datasets

The foundational catalogues of pre-existing bioisosteric pairings are in the comprehensive review articles listed above, and others that exist within the wider literature.[97,154,158,168,196–201] These have the advantage of being carefully curated by scientific authors, and the examples chosen within them are usually well-characterised within the published literature. Furthermore, these are frequently annotated with the underlying scientific rationale, thereby clarifying the contexts in which each pairing is appropriate and supporting informed selection of bioisosteres.

Originally designed to provide a means of systematically searching the reviewed bioisosteric literature, BIOSTER is a proprietary, manually-curated database of compound pairings that have been documented in published literature as being biologically interchangeable.[203] Based initially on review articles like those described above, the scope was expanded to include a wider range of the published literature, and pairings are extracted from over 100 journals and periodicals, including pesticides and agricultural chemicals. It contains nearly 30 000 entries, and each one is annotated with the target, and relevant literature references. It exists, however, behind a proprietary paywall and is reliant on manual curation and updating, which makes it susceptible to omissions; some relevant pairings may be missed if not captured in the curated sources. Moreover, as a retrospective resource, it cannot propose novel bioisosteric substitutions that have not yet been reported in the literature.

1.4.1.1 Mining the PDB and ChEMBL

The Protein Data Bank (PDB) is an online, open-access repository of experimentally determined three-dimensional structures of biological macromolecules, and at the time of writing contains 233 605 macromolecular structures, of which 45 931 are bound to small molecule

ligands (this figure includes inorganic ions, natural cofactors, and synthetic ligands).[204,205] This dataset presents opportunities for the discovery of bioisosteric pairings, as the structural data available allows for the characterisation of the steric and electronic parameters of the binding site, and the binding pose of the ligand within it. Several tools have been developed in the literature that take advantage of the PDB as a source of bioisosteric data, and some these are outlined here.

Kennewell *et al.* described in 2006 a method for identifying bioisosteres by aligning protein–ligand complexes from the PDB based on shared protein sequences, extracting and fragmenting the ligands, and comparing the fragments to identify structurally and spatially overlapping substructures.[206] Fragment pairs are scored based on a computational approximation of molecular volume overlap using atom-to-atom distances, and only those surpassing a defined similarity threshold, and passing chemical filters, are retained as candidate bioisosteres.[207] Their method identified within 10 minutes a set of diverse, spatially overlapping fragment pairs from PDB ligand sets highlighting their potential as target-specific bioisosteres for lead optimisation. The source code was not released with the paper.

The key representation of interactions in pockets (KRIPO) method described by Wood *et al.* in 2012 generates structure-based pharmacophore fingerprints from PDB protein–ligand complexes by fragmenting ligands, defining local binding sites, and encoding intermolecular interaction features (e.g. hydrogen bonding, electrostatics, π-stacking) into 'fuzzified' 2- to 4-point fingerprints, which are then used to assess binding site similarity.[208] These optimised fingerprints were used to create a database of nearly 300 000 local binding site fingerprints, with the authors demonstrating that the fingerprints were able to distinguish between similar and dissimilar binding sites across different protein families. Ligands that bind in sites with similar 'fuzzy' fingerprints are proposed to be bioisosteric; crucially this determination is

solely based on the ligand binding site, and takes no account of the nature of the protein or its global structure. The database was not released with the paper, however.

Two graph-based methods that identify bioisosteric pairings from PDB data without necessitating pairwise similarity calculations between ligands or binding sites are sc-PDB-Frag and FragVLib.[209,210] The sc-PDB-Frag database is built by fragmenting ligands from 8077 curated protein–ligand complexes.[209] Ligands were first fragmented using one of two literature protocols (HOME and RECAP, based on those described in references [211] and [212] respectively), and fragments that formed at least four distinct interactions with their protein environment were retained.[211,212] Each fragment was annotated with ligand and target metadata, and paired with computed interaction fingerprints and graphs, proposing that ligands which shared similar interaction graphs and fingerprints with their respective targets could be bioisosteric.[209] Similar to this, FragVLib found bioisosteric fragments by comparing the shapes and chemical features of protein–ligand binding sites by representing the interaction points between protein and ligand as a graph (without initially fragmenting ligands like in sc-PDB-Frag), then identifying similar patterns with sub-graph matching from a database extracted from the PDB. Any ligands identified with similar interactions to a query are then mapped into the binding site of the query protein, creating a virtual library of fragments (FragVLib) tailored to the shape and electrostatic environment of the query protein. Both of these methods and their respective databases were released as open-source projects, and are accessible to users for free via a web-browser.

Like sc-PDB-Frag and FragVLib, the Base of Bioisosterically Exchangeable Replacements (BoBER) leverages structural data from the PDB to identify bioisosteric replacements, but using binding site superposition (via the ProBiS algorithm described by Konc *et al.*) to define structural similarity between pockets, followed by fragment matching based on spatial

overlap rather than pharmacophoric or interaction-based features.[213,214] The binding sites of proteins extracted from the PDB were aligned with ProBiS, and any that had a similarity above a threshold (measured as a Z-score $\geq$ 2) had their bound ligands superimposed, fragmented, and compared based on spatial overlap. These results were compiled into a searchable database on the internet, which is freely accessible through a user-interface to academic users.[214]

A more recent example is Zhang *et al.*'s BioIsoIdentifier, which aims to identify bioisosteric replacements by searching PDB ligands containing a user-defined substructure, superimposing homologous proteins, and extracting overlapping fragments based on 3D shape and electrostatic similarity using the ShaEP algorithm by Vainio *et al.*[215,216] Extracted fragments are clustered by molecular similarity, and results are visualised through a freely-accessible interactive web interface.

The ChEMBL database is another key source of biological information. ChEMBL is a manually curated, open-access database of bioactive molecules with drug-like properties, containing detailed information on their chemical structures, biological activities, and molecular targets; at the time of writing ChEMBL 35 has nearly 2.5 million individual compounds annotated with data from over 1.7 million biological assays.[194,217,218] The SwissBioisostere tool makes use of ChEMBL to identify bioisosteric pairings.[219,220] This is achieved using a matched molecular pairs analysis of the ligands in ChEMBL annotated with biological information.

Matched molecular pairs (MMP) analysis is a powerful technique for quantifying the effects of structural changes, where pairs of compounds differing by only a single, well-defined chemical transformation are identified and compared, enabling direct evaluation of property

changes such as potency, solubility, or selectivity.[221] An MMP is illustrated in Figure 1.22 panel A, extracted from Pastor *et al.*'s investigations into PIM1 kinase inhibitors.[222] **1.21** and **1.22** can be considered as a matched molecular pair as they differ only in their core scaffold (highlighted in bold), noting that a rearrangement of the heteroatoms has improved the binding affinity by a factor of 2. There are several algorithms for identifying MMPs in a dataset, of which the most efficient is that described by Hussain and Rea, and is illustrated in Figure 1.22 panel B.[221,223] Exocyclic bonds are cut (illustrated by dashed red lines) in each molecule, and the resulting fragments are stored in a dictionary as a key-value pair. If later molecules in the dataset share fragments with a previous molecule (as **1.22** would), the variable fragment is appended to the entry associated with the shared, constant portions. It is thus possible to identify, by examining the values for each key, the precise change that defines the MMP, which in this case is classified as offering an improvement in PIM1 kinase inhibitory activity.

SwissBioisostere was constructed initially by mining ChEMBL 13 for compound pairs tested on the same target with well-defined bioactivity data (IC_{50}, EC_{50}, K_i, or K_d).[219,220] After curation, including standardisation, filtering by target confidence, and removal of ambiguous activity values, MMPs were identified as described above, fragmenting on exocyclic bonds. Only replacements meeting criteria on size, stereochemistry, and topological similarity were retained. The resulting bioisosteric transformations, annotated with attachment context and biological data, were made freely available on the internet, and are searchable through a web interface. It was updated in 2022 to include data from ChEMBL 28, and contains over 25 million bioisosteric replacements, annotated with references and a measure of the biological effect of the replacement.[220]

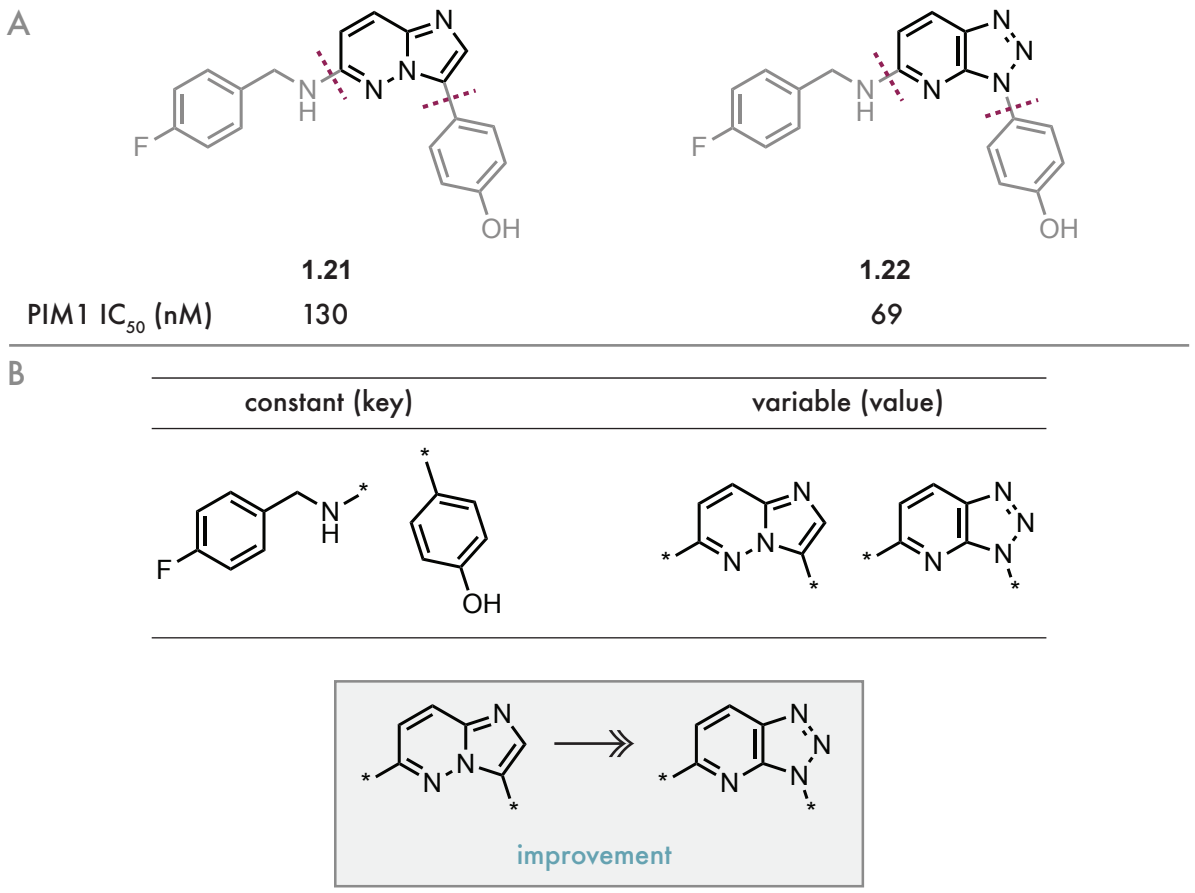


Figure 1.22. An illustration of matched molecular pairs. **A** | Matched molecular pairs from a dataset of PIM1 kinase inhibitors as part of an oncology drug discovery campaign.[222] The well-defined structural change in this case is the variable aromatic scaffold, and the constant regions the side-chains. The dashed lines indicate where the cuts are made to separate the constant portions from the variable core. **B** | The Hussain-Rea algorithm for identifying MMPs, as implemented in the creation of SwissBioisostere.

These encyclopaedic tools of previously characterised data are clearly useful (at the time of the 2022 update SwissBioisostere reported a total of 45 900 unique users), but they rely on molecules that have already been designed, synthesised, and assayed, and as such are limited in their ability to expand the scope of bioisosteric replacements available to medicinal chemistry campaigns. Tools for the proposal of truly novel bioisosteric replacements rely either on screening of virtual chemical libraries or generative statistical methods, and are outlined in the next section.

1.4.2 Virtual Screening Approaches

The techniques that are able to identify wholly novel bioisosteric pairings can be broadly classified as those which screen virtual libraries for similarity to a query ligand, and those that use statistical methods (including machine learning) to propose novel structures based on a training dataset. This section will outline some of the *in silico* techniques previously developed for screening virtual libraries, and give a brief overview of several open-source virtual libraries available for screening. A detailed description of machine learning and statistical approaches to bioisostere discovery is outside of the scope of this thesis, and thus readers are referred to references [46, 224–226] for recent reviews and examples.

1.4.2.1 Virtual Screening Techniques

Virtual screening techniques are traditionally divided into those which are ligand-based and those which are structure-based. Structure-based virtual screening exploits information about the 3D structure of the target protein to identify molecules which are predicted to bind strongly to it (a key example of this being molecular docking), and although a review of these techniques is outside of the scope of this thesis, there are many comprehensive reviews of this important branch of computational drug discovery.[227–231] Ligand-based virtual screening involves comparing a known, active query ligand against a database of molecules to identify those that are similar by some metric; for example molecular volume, electrostatics, or a chosen physicochemical property.[228]

Ligand-based screening can be further categorised into 2D screening, involving the characterisation of ligands by 2D descriptors (for example the fingerprints discussed here) indicating the presence or absence of a certain molecular feature (Figure 1.23), and 3D screening, which

involves alignment of the 3D ligand structures followed by scoring.[i232] Fingerprint methods are computationally faster, as there is no need to align ligands or explore conformational space and can be compared using efficient, bit-wise similarity calculations, but they are abstractions, and as such there is information loss associated with them.[232] Longer fingerprints encode more information, and have been shown in general to perform better in virtual screening than shorter fingerprints, however despite the associated abstractions, Venkatraman *et al.* demonstrated that 2D fingerprint searching can outperform certain 3D shape-based methods under some circumstances.[233,234] Fingerprints have been reviewed extensively in reference [232], but a few commonly used examples include the MACCS substructure-based fingerprint (available as both a 166-bit and a 960-bit version), the 881-bit PubChem structural fingerprint, Daylight's 2048-bit topological fingerprint, extended-connectivity fingerprints (ECFPs) based on the environment of each atom about a fixed radius, and hybrid fingerprints such as the 171-bit MP-MFP.[232,235–238]

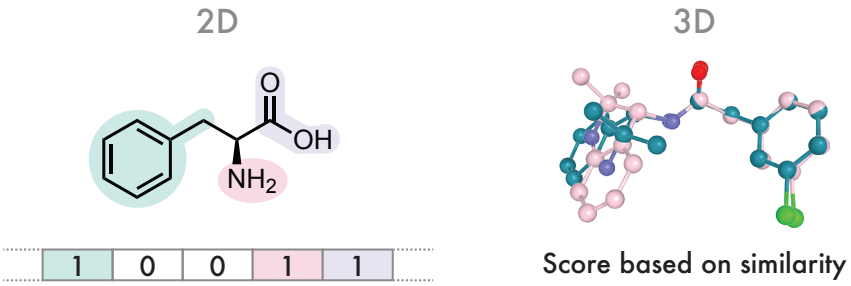


Figure 1.23. A comparison of 2D vs 3D virtual screening. 2D substructure-based fingerprinting constructs a vector representation of a molecule, where each element of the vector corresponds to the presence or absence of a particular defined molecular feature (in the example above the presence of a benzyl, amine, and acid). These are then compared in a bit-wise fashion to other fingerprints, whereas 3D methods rely on alignment, superposition, and scoring, and are thus inherently more conformationally dependent.

[i]Fingerprints based on the 3D structure of ligands do exist, but the comparison methods do not involve ligand superposition, and as such fall here under the general umbrella of 2D methods.

3D virtual screening methodologies have been the subject of thorough reviews by Langdon *et al.* and Oliveira *et al.*, and differ from the fingerprint-based approaches outlined above due to the need for information about the 3D structure of the ligands in the library.[224,239] A seminal work in this area was that of Lauri and Bartlett, with their CAVEAT program in 1994.[240] CAVEAT introduced a structure-based approach for identifying bioisosteres by searching 3D molecular databases for fragments that could position key functional groups (exit-vectors) in the same spatial arrangement as those in a known active compound. Rather than relying on overall molecular similarity, the method focused on preserving geometric and chemical features critical for target binding, and thus is an early example of applying 3D geometry to virtual screening and scaffold hopping.

The Rapid Overlay of Chemical Structures (ROCS) is another 3D virtual screening tool, described in 2005 by Rush *et al.* that identifies molecules with similar shape to a query compound by aligning their molecular volumes and maximising their spatial overlap.[241] Instead of relying on atom types or connectivity, ROCS models each molecule as a set of smooth Gaussian functions centered on atomic positions, allowing for efficient volume-based comparisons. The degree of shape similarity is quantified based on the overlapping volumes, enabling the discovery of structurally diverse molecules that share similar 3D shape and thus may exhibit similar biological activity. ROCS is efficient, screening around 800 molecules a second, but is a proprietary software, and does not include any measure of electrostatic complementarity. ROCS was used as part of Pfizer's in-house ligand virtual screening tool NEAT, described in 2012, which uses ROCS and an electrostatic similarity measure to search through a custom-built virtual library of molecules derived from the GDB11 database, however neither the code nor the library were distributed publicly.[242,243]

ShaEP, designed by Vainio *et al.*, combines 3D shape overlap and electrostatic potential (ESP) similarity by comparing field-graph representations of molecules.[216]. Field-graphs are constructed by placing Gaussian-weighted vertices around a molecule to represent shape and labelling these with a measure of the ESP at each position, then using subgraph isomorphism matching to align molecules based on these features. The final similarity score is a weighted average of the shape overlap and electrostatic field similarity, allowing accurate comparison of molecules with similar binding properties but diverse structures. It is freely distributed, and is well-benchmarked against the directory of useful decoys (DUD).[244]

Also using Gaussian representations of molecular shape are Pharao, which detects pharmacophore features from 3D structures and aligns them by maximising the overlap of their Gaussian volumes, enabling rapid and flexible comparison of molecular interaction patterns, and LIGSIFT, which aligns molecules by maximising the overlap of atom-based shape and chemical densities using a cost-optimised matching algorithm and Monte Carlo refinement, producing size-independent similarity scores for virtual screening.[245,246] Despite being efficient, these methods do not provide the user with atomic-resolution alignments. LSAlign is an open-source tool designed by Hu *et al.* which performs both rigid and flexible 3D ligand alignments by optimising structural and chemical similarity scores (LS-score and PC-score), and evaluates alignments using a statistically normalised significance model.[247] Crucially LSAlign is able to return alignments of highest similarity to the user, enabling later-stage functionalisation and the integration of the aligned results into wider virtual drug-discovery pipelines.

Recognising that ligand alignment is often a bottleneck in ligand-based virtual screening workflows, Ultrafast Shape Recognition (USR) algorithms were developed in the laboratory of Graham Richards at Oxford to circumvent this.[248–250] These methods avoid explicit molec-

ular superposition by encoding shape using distributions of atomic distances from strategically chosen centroids, enabling rapid and alignment-free comparison of molecular shapes across large databases.[249] ElectroShape incorporated partial atomic charge as a fourth spatial dimension, thereby allowing a simultaneous comparison of ligands by shape and electronic distribution, and was shown to improve enrichments over shape-only USR methods.[250]

1.4.2.2 Virtual Libraries

In order to identify bioisosteres, the tools outlined above require virtual libraries of compounds to compare to the specified query ligands. Many pharmaceutical companies have their own in-house virtual compound collections, for example the Pfizer collection used in the NEAT screening outlined above, which can either be based on physical libraries of compounds managed by the company, or virtual libraries built using enumeration strategies.[242,251] It is possible also to use the catalogues of molecules available from commercial suppliers, which often extend into regions of tangible but unexplored chemical space enabled by combinatorial chemistry techniques on existing building block libraries.[251,252] Examples of these synthetically feasible libraries include the MCule database of over 100 million compounds, WuXi's Galaxi's library of over a billion compounds, and the Enamine REAL library of over 9.6 billion molecules.[253–255] Ruddigkeit *et al.* published the chemical universe database GDB-17 in 2012, a comprehensive enumeration of 166 billion stable, synthetically accessible organic molecules containing up to 17 atoms of C, N, O, S, and halogens.[243]

These libraries cover a large region of chemical space, and thus it is highly likely that within them lie many new bioisosteres of common medicinal chemistry moieties, however their very large size at the moment precludes their searching on feasible timescales, although advances in AI and high-performance computing are likely to see this change in the future.[225]

The VEHICLe database, published by Pitt *et al.* in 2009, is a library of 24 867 aromatic heterocycles representing a complete enumeration of a section of chemical space defined by the following parameters:

- five-membered or six-membered rings, and bicyclic combinations thereof;
- containing only C, H, O, N, S;
- only exocyclic carbonyls (i.e. all molecules exist in their keto tautomers);
- electrically neutral; and
- all obeying Hückel's $4n + 2$ rule of aromaticity.

This method of constructing VEHICLe means that all possible annular tautomers of heterocycles capable of tautomerism are represented as separate structures within the database. However, because every ring is required to be aromatic, tautomers that break aromaticity are not included. For example, in the case of phenol (see Figure 1.5), the corresponding keto tautomer disrupts the aromatic π-system and is therefore not included. Hydroxyl substituents are also not represented in the database, so the enol form is likewise absent. Many heterocycles exist with the potential for lactam-lactim tautomerism, however the construction requirement for carbonyls rather than hydroxyl groups mean that only lactam tautomers are included in the database.

As this library includes all possible heterocycles that meet these criteria, it therefore contains both commonly encountered heterocycles in chemistry and those that have never previously been synthesised. It is proposed that within this library of molecules lies previously unsynthesised heterocycles that are both accessible through contemporary synthetic methodologies

and bioactive.[256] The distinct possibility exists therefore that these molecules are bioisosteres of commonly used heterocycles in medicinal chemistry, and their inclusion in the arsenal of chemical functionality available to the medicinal chemist could allow access to chemical space with novel, beneficial physicochemical properties.

The authors suggested at the time of publication in 2009 that although they estimated over 3000 of the 24 867 VEHICLe heterocycles to be accessible with contemporary synthetic methodologies, only 1701 of them had been documented in the Beilstein database (*vide infra*).[256] To gain a crude estimate of the potential for VEHICLe to contain novel, synthetically feasible heterocycles, as part of this thesis substructure searches were run of all VEHICLe heterocycles through the ChEMBL 31 database of 2 331 700 bioactive molecules in the manually curated database.[217] This includes molecules published and assayed in the literature, those involved in clinical trials, and those in FDA-approved drugs. This located 1618 of the VEHICLe heterocycles in ChEMBL31, and only 489 of these in FDA-approved drugs. This information is displayed in the outer ring of Figure 1.24. Although this figure is similar to the 1701 quoted by the original authors, the discrepancy is likely due to the differing nature of the databases used to conduct the analysis. Whereas the search documented here was through ChEMBL 31, which is a database specifically geared towards bioactive molecules, the original authors searched through the Beilstein database of chemical compounds.[217,256] The Beilstein database, owned since 2009 by the publisher Elsevier and accessed as Reaxys®, is a comprehensive dataset of over 6 million molecules extracted from all chemical literature published since 1771, and therefore includes molecules and structures that have never been assayed or tested in the biological literature. The discrepancy between Pitt *et al.*'s original number and the 1618 reported here is therefore likely to exist due to an incomplete overlap of the two databases searched, and the likelihood that more heterocycles have been documented

in the 16 years since the original publication. Despite this minor discrepancy, the conclusion of this analysis is in keeping with those reached by Pitt *et al.*: that a large number of heterocycles within this database remain unexplored but synthetically accessible. As the goal of this work is to uncover novel molecules primarily for use in biological applications, molecules that have been reported in synthetic literature but have never been included within small molecule drug discovery projects sit under the general umbrella of 'biologically unexplored compounds', and therefore count as unsynthesised for the purposes of this analysis.

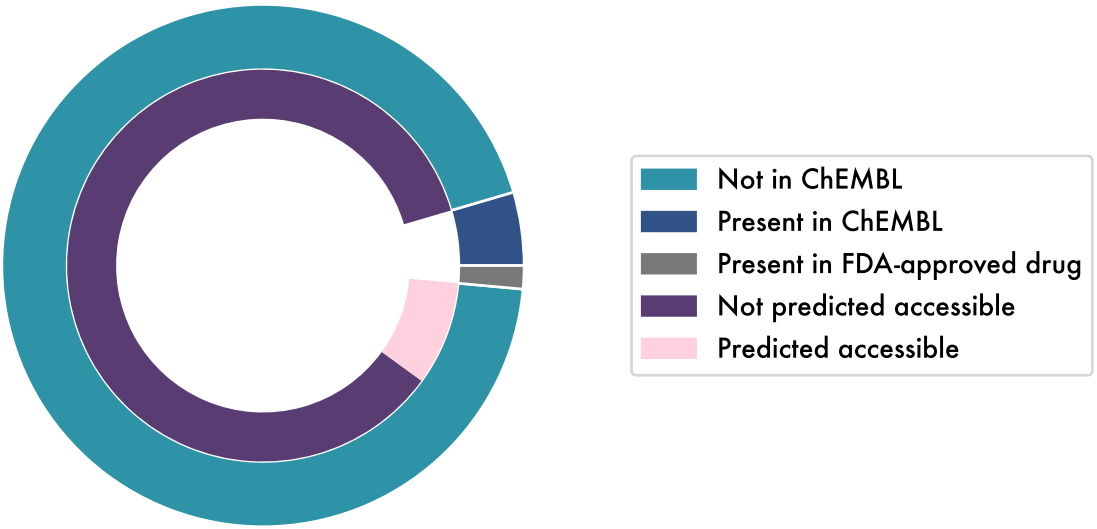


Figure 1.24. An estimate of the synthetic accessibility of VEHICLe using the SAScore as a measure of synthetic accessibility.

To gain an estimate of the proportion of the unsynthesised molecules within VEHICLe that may be accessible with contemporary synthetic methodology, a synthetic accessibility analysis was performed using the SAScore metric by Ertl and Schuffenhauer.[257] The SAScore uses a machine-learned model trained on over 900 000 molecules from the PubChem database to

determine a synthetic accessibility score based on the occurrence within a molecule of 'fragments' that are known to be synthesisable, and penalises the existence within the molecule of certain complexity flags. It awards a molecule a score out of 10, with scores > 6 deemed to be difficult to synthesise, and those ≤ 2 synthetically facile. These scores were calculated for each VEHICLe heterocycle, using the RDKit implementation of SAScore in Python. The distribution of scores for the heterocycles that feature in FDA-approved drugs, and therefore are both synthesisable and biologically active, was examined and the upper quartile of their scores was taken as an threshold value to demarcate those previously unsynthesised heterocycles that could be considered to be accessible. This information is displayed in the inner ring of Figure 1.24. As a result of this analysis, it is estimated that at least 2130 of these previously unsynthesised molecules are synthesisable. This is a conservative estimate, and it is likely that a greater number of the unsynthesised heterocycles in VEHICLe are accessible.

This analysis highlights the unexplored potential of the VEHICLe library as a source of novel, synthesisable heterocycles. With over 2000 previously unsynthesised and synthetically accessible structures, VEHICLe represents a tractable and comprehensive chemical space in which novel bioisosteres of medicinally relevant heterocycles are likely to be found.

1.5 Thesis Aims

Given the ubiquity of aromatic heterocycles in small molecule drugs, and the significant physicochemical and ADMET improvements to lead molecules that can be gained by the use of aromatic heterocycles as bioisosteres, it is clear that expanding the region of aromatic heterocyclic bioisosteric space that is accessible to medicinal chemistry can only be beneficial to the process of creating effective new drugs. The objective of the work presented in this the-

sis is to develop computational tools for proposing novel aromatic heterocyclic bioisosteres of common heterocycles in medicinal chemistry using ligand-based virtual screening techniques on curated compound libraries. The inclusion of previously unsynthesised heterocycles that are predicted to be synthetically accessible allows for the discovery of completely novel aromatic heterocyclic bioisosteres with new physicochemical and ADMET profiles, and has the potential to direct heterocyclic synthetic methodology efforts towards regions of chemical space that are more likely to be biologically active.

The first section of this thesis (Chapters 2, 3, and 4) describes initial efforts to search VEHICLe using the ShaEP methodology by Vainio *et al.*, the expansion of VEHICLe to include ring substituents through the creation of the MoBiVic library, and the development of a new, vector-based alignment and scoring method for searching this library of functionalised heterocycles, ultimately resulting in the Heterocycle Isostere Explorer (HCIE), an open-source Python package for discovering novel aromatic heterocyclic bioisosteres.[216] The subsequent sections (Chapters 5 and 6) describe experimental attempts to synthesise novel heterocycles from VEHICLe using disconnections predicted by a machine-learning retrosynthesis model developed in the Brennan and Duarte groups, and the application of the methodologies described in the first section to ongoing medicinal chemistry projects within the Brennan group.

2 Development of the Initial Implementation

> Bioisosteres are compounds or groups that possess *near-equal molecular shapes* and volumes, *approximately the same distribution of electrons*, and which exhibit similar physical properties such as hydrophobicity. Bioisosteric compounds affect the same biochemically associated systems as agonists or antagonists and thereby produce *biological properties that are related to each other*.
>
> Alfred Burger
>
> Isosterism and Bioisosterism in Drug Design (1991)

Within Alfred Burger's 1991 treatise on isosterism and bioisosterism in drug design, he proposes a definition of bioisosterism that clearly links compounds with similar biological properties to similarities in their molecular shapes and electronic distributions.[158] Encapsulated within this simple definition is a good deal of complexity, not least in the understanding of what constitutes *biological properties that are related to each other*, but its appeal lies

in the simplicity with which Burger establishes the criteria for bioisosterism. By directly linking similarity in molecular shape or volume and electrostatic potential to similarities in biological activity, this definition (summarised in Figure 2.1) provides a set of criteria which allow the discovery of novel bioisosteres.[158]

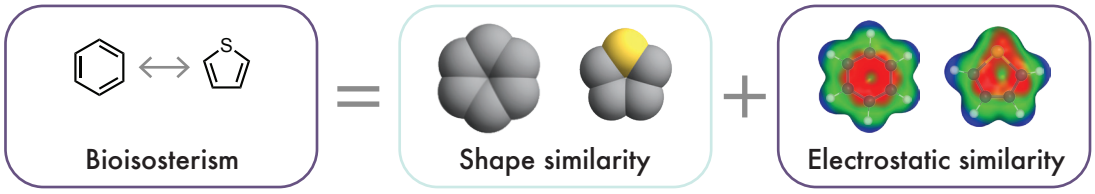


Figure 2.1. Bioisosterism, based on a definition proposed by Alfred Burger in 1991.

Taking this definition as a starting point, this chapter outlines the development of the first iteration of the Heterocycle Isostere Explorer (HCIE): a computational tool for searching a virtual database of aromatic heterocycles to identify potentially bioisosteric molecules of a user-specified query. By screening a virtual library for shape and electrostatic similarity to the query, heterocycles similar by Burger's definition can be identified.

2.1 Methodology

The strategy described in this chapter for proposing novel bioisosteric pairings is to search a large, virtual library of aromatic heterocycles for those similar in shape and electrostatic surface potential to a user-defined query molecule. In order to appeal to scientists with differing levels of computational experience and maximise the probability of returning novel bioisosteric pairings, it must meet several criteria:

1. Be easy to install and use.

2. Have a clear and accessible mechanism for specifying the structure of the query molecule.

3. Search as complete an enumeration of chemical space as possible, thereby ensuring maximal structural and electrostatic diversity in the candidate molecules.

4. Be computationally efficient, returning ranked results within a reasonable time frame for a large library of molecules.

5. Provide a simple means of visualising the molecular structures of the outputs, ideally in the alignment of highest similarity to aid subsequent compound design.

The methodology described herein compares a user-defined query molecule against every heterocycle in the VEHICLe database, using the ShaEP algorithm designed and implemented by Vainio *et al.* in 2009.[216,256] The selection of VEHICLe as a virtual database and its potential for uncovering novel, previously unsynthesised bioactive aromatic heterocycles is discussed. The pre-processing of the VEHICLe and query ligands, including geometry optimisation, partial charge calculation, and the extraction of atomic connectivity information in preparation for searching and alignment is described, alongside a simple explanation of the underlying methodology of the ShaEP algorithm.

2.1.1 VEHICLe as a Searchable Virtual Library

The VEHICLe database of 24 847 aromatic heterocycles, assembled by Pitt *et al.* in 2009 and outlined in Section 1.4.2.2, was chosen as the virtual library for the initial implementation.[256] An overview of the key physicochemical properties of the VEHICLe heterocycles is shown in Figure 3.5.

The inclusion within VEHICLe of molecules that have not been synthesised comes with several advantages. Exhaustively searching as diverse a range of heterocyclic chemical

space as possible maximises the chances of discovering unconventional but viable bioisosteres that have yet to be characterised. A wider pool of bioisosteric candidates also maximises the chances of finding molecules with preferential physicochemical or ADMET properties. Searching a smaller subset, perhaps selected based on those previously synthesised or those appearing already in bioactive molecules, risks biasing results towards these known molecules.

Another key advantage of using VEHICLe as the initial searchable library is that the database is open-source, and thus all of the molecules are available to freely download as SMILES strings. This makes distribution of the searching package straightforward, as the database is freely accessible to all users, and saves on overhead time as there is no need to reproduce the dataset from the original literature description.

2.1.2 ShaEP as a Searching Algorithm

Having selected VEHICLe as the searchable database, a method for aligning and scoring the VEHICLe heterocycles against a query molecule was needed. This method would need to be freely distributed, quick to search through a large library, and return structural information (*i.e.* the alignments of highest similarity) to enable the use of the results in subsequent compound design. Its searching and scoring algorithm should also be compatible with the hypothesis illustrated in Figure 2.1, scoring ligands based on ESP and volumetric similarity.

As previously outlined in Section 1.4.2.1, there exist several software tools that might be useful for searching VEHICLe according to the criteria described above. The USR algorithm ElectroShape incorporates atomic partial charge as a fourth spatial dimension, thereby allowing a very efficient search of a virtual library by both shape and electrostatic similarity, however it does not superimpose or align the ligands, and therefore does not return the neces-

sary structural information to enable further compound elaboration.[250] Pfizer's tool NEAT, based on the proprietary ROCS algorithm, does align and score ligands based on shape and ESP similarity, however it is not freely distributed, and its source code is proprietary. Of the tools described in Section 1.4.2.1, the ShaEP algorithm (named as an amalgamation of shape and ESP) described by Vainio *et al.* in 2009 fulfils the greatest number of criteria outlined above.

ShaEP efficiently searches large libraries of compounds, comparing each library molecule to a query molecule accounting for both the similarity in the atomic positions (shape) and in the distribution of electrons (ESP).[216] Although its source code is proprietary, a binary executable is freely distributed and it is well documented. The algorithm had been benchmarked and tested in its original publication, and was shown to take a median time of 125 ms per structure, so would thus screen the entire VEHICLe library in approximately one hour[a]. Furthermore ShaEP outputs the highest scoring alignment for each heterocycle as an `.sdf` file, so the alignments can be visualised using any freely-distributed molecular visualisation software. ShaEP was therefore selected as the alignment and scoring algorithm for the initial implementation of the HCIE searching tool.

ShaEP requires that the query molecule and the all of the library probe molecules be provided to it with a pre-determined 3D geometry and electrostatic information in the form of partial charges, therefore the library and query molecules require a degree of pre-processing before alignment and scoring. The overall workflow employed in this implementation and described throughout this chapter is summarised in Figure 2.2.

[a]As the VEHICLe molecules are smaller than those in the benchmarking set, this search time is likely to be significantly reduced in this implementation.

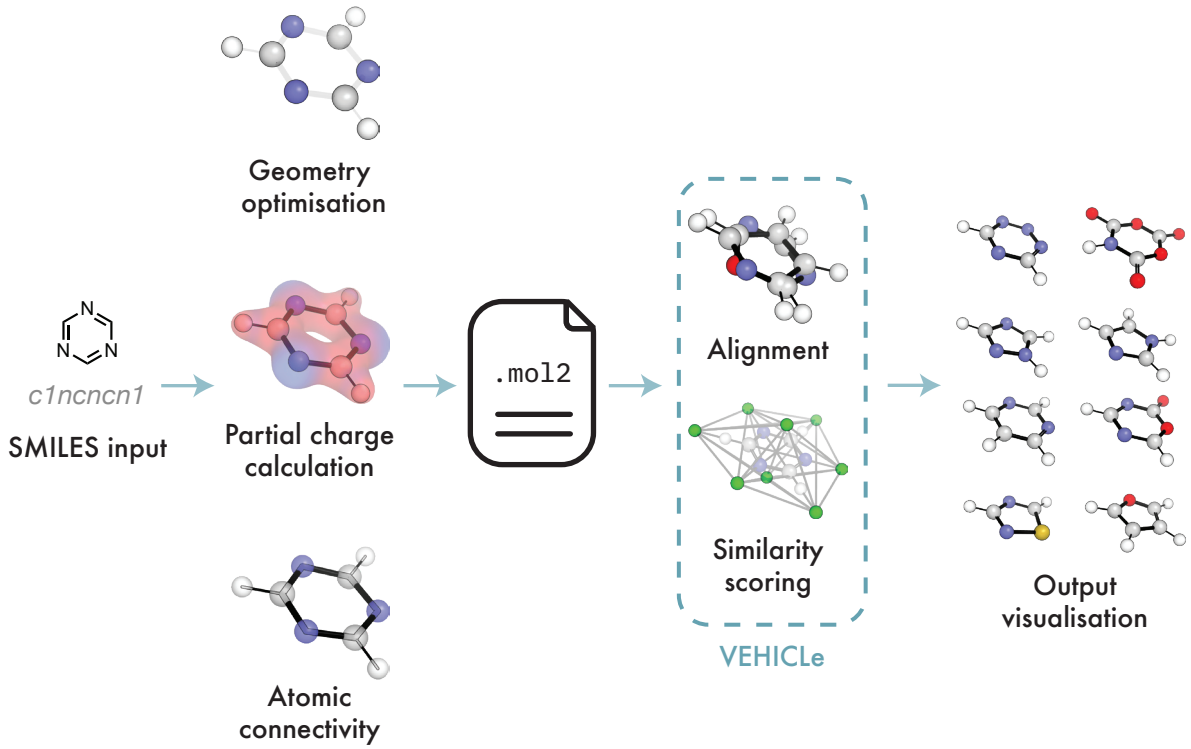


Figure 2.2. Workflow for the first-generation HCIE algorithm. The alignment and similarity scoring steps are carried out by ShaEP.

Initially the molecules are specified as a SMILES string by the user. SMILES strings are string-based representations of molecules that encode the molecular connectivity and information about the electronic nature of the atoms and bonds in the molecule.[258,259] They are a universal, easily understood, and computer-readable molecular format that can readily be generated from graph-based molecular representations using commonplace chemical drawing software (for example ChemDraw™ or MarvinSketch™), and are lightweight and memory-efficient. Once the user has described the query molecule by means of a SMILES string, a digital representation of the molecule is created which enables the optimisation of the molecular geometry and the calculation of atomic partial charges using the autodE software developed in the Duarte Group (*vide infra*).[260] Information about the atomic connectivity is also extracted from this digital representation of the molecule; specifically which atoms and bonds are aromatic. All of this information is then compiled into a TRIPOS `.mol2`

file, which is then used as the input for a ShaEP search with the VEHICLe database as the screening library. Each VEHICLe heterocycle is optimised and partial charges and connectivity calculated, and these are stored as a library `.mol2` file. The query molecule is then sequentially aligned to and scored against each heterocycle, and the scores of the top 20 heterocycles and their alignments are returned to the user. The scores of all the heterocycles in the library are returned to the user in a separate `.txt` file.

This section describes in detail the various stages outlined in Figure 2.2, beginning with the instantiation and pre-processing of the query molecule, including the geometry optimisation and calculation of the partial charges. An overview of the ShaEP alignment process is then given, before the output and visualisation of results is described.

2.1.3 Pre-Processing

Following the user specifying the query molecule by a SMILES string, an autodE `Molecule` object is created from the SMILES string. autodE is an open-source Python package developed in the Duarte group which acts in this context as a SMILES parser and convenient wrapper for the electronic structure packages used here for geometry optimisation and partial charge calculation. It was chosen for its convenient Python interface, compatibility with RDKit, and easy access to both DFT and wavefunction based methods, and is well documented and tested. This autodE `Molecule` object is then straightforward to optimise and calculate charges for, using the object-oriented implementations of various electronic structure packages wrapped in autodE.

Once the computational representation of the molecule has been created, each query (and each probe in the library) must undergo a geometry optimisation, and then partial charges

calculated on this optimised geometry before any searching can take place. A brief description of the necessity of these steps is given below, followed by a description of the xTB-derived geometries and charges employed in this initial approach.[261]

2.1.3.1 Geometry Optimisation

In the physical world molecules exist as three-dimensional structures, thus the two-dimensional representations such as the skeletal formulae commonly drawn by chemists or that encoded by a SMILES string do not accurately reflect the physical shape of a molecule. In order to appropriately capture the steric similarity between two ligands, it is therefore necessary to approximate computationally the 3D geometry of each molecule as close as reasonably possible to its true physical conformation.

The geometry (or structure) of a molecule is specified by the coordinates of each of the atoms that constitute it. Within the Born-Oppenheimer approximation each atomic configuration has a ground-state energy found by solving the Schrödinger equation. The equilibrium geometry is the atomic arrangement that minimises this energy. This is illustrated in Figure 2.3, where the surface shown represents the energy of a particular arrangement of atoms given by arbitrary bond lengths one and two. The red circle indicates the equilibrium geometry, which corresponds to the bond lengths that minimise the energy of the molecule. In order to fairly compare ligands, it is necessary that each of their geometries is optimised such that it sits as close as possible to the geometry which minimises the energy. If the geometry of one ligand is not sufficiently close to the minimum (for example that indicated by the yellow circle), then physically realistic representations of molecules are not being compared, and the results will not be meaningful.

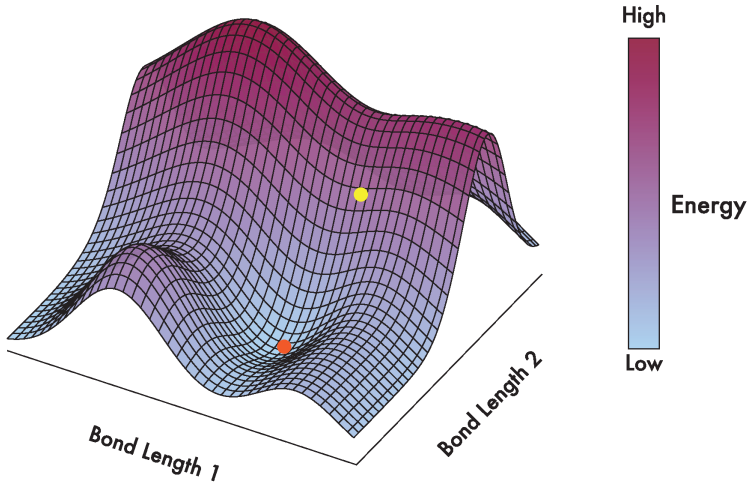


Figure 2.3. The geometry dependence of the energy of a hypothetical molecule defined by two bond lengths 1 and 2. The red circle indicates a molecule close to its equilibrium geometry, and the yellow circle a molecule far from its equilibrium geometry.

Minimising the energy of a molecule directly from the Schrödinger equation is very computationally expensive, and as such there have been developed a large number of techniques that aim to approach the minimum energy geometry in an efficient but computationally cheap manner.[262] An in-depth discussion of these is outside of the scope of this thesis. As the molecules in the VEHICLe database are all unsubstituted aromatic heterocycles, the conformational space available to them is limited, and their PES is likely to be relatively flat, thus it is probable that a simple optimisation algorithm will be sufficient to ensure they each take a reasonable geometry before alignment and scoring.

2.1.3.2 Partial Charges

In order to score molecules based on their electrostatic similarity, it is necessary to approximate the electrostatics of the molecule computationally. Electrostatic surface potentials represent the electrostatic potential energy experienced by moving a unit positive charge about a defined surface around a molecule.[263] As the distribution of electrons about het-

eronuclear molecules is anisotropic in nature, the value of this potential will vary depending on whether a particular region of the molecule is electron-rich or electron-poor. Hydrogen-bonding, salt-bridge formation, sigma-hole, and pi-stacking interactions are inherently electrostatic in nature, therefore much of the necessary information to determine the electrostatic complementarity of a ligand with its target is encapsulated within its ESP.[264]

To accurately represent these, it is necessary to map the distribution of charge at an atomic level within the molecule. The distribution of charge, much like the geometry described earlier, is dependent on the nature and connectivity of the nuclei in a molecule, which thus determines the number and distribution of electrons.[b] It is common in computational modelling of small molecule electrostatics to use the partial charge on each atom as a way of approximating the molecular electron distribution. Although this is not a physical observable, it is a useful model for determining the relative distributions of electrons, and can be calculated in many different ways, each with differing levels of abstraction and complexity.[265,266] Fundamentally this involves partitioning the continuous electron density distribution of the molecule into atomic contributions. These atomic partial charges, which for the molecules in the searchable database will always take non-integer values between -1 and 1, represent the electron density that can be considered to be 'more localised' around that particular atom.

The concept of partial atomic charges is demonstrated simply in Figure 2.4. This figure shows that for the highly symmetrical molecule benzene, the aromatic carbons each carry a partial charge of -0.06, and the hydrogens +0.06. This is to be expected, as carbon is more electronegative than hydrogen and therefore draws a greater share of the electron density in the bond towards itself. It should be noted that sum of the partial charges is equal to the overall charge of the molecule, which for all VEHICLe molecules will be 0. In the

[b] For a neutral molecule.

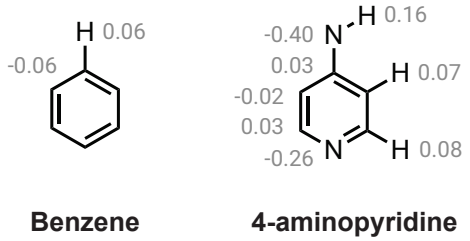


Figure 2.4. An example of the concept of atomic partial charges in two simple molecules. The charges displayed are Gasteiger charges calculated using the method of electronegativity equalisation; this is an empirical gas phase model and does not account for solvation. Symmetrically equivalent hydrogens are omitted for clarity.

more electronically complex 4-aminopyridine the pyridine nitrogen carries a large $\delta-$ charge, whereas the ortho- and para-carbons are slightly $\delta+$ as would be expected from a simple valence bond and resonance argument.

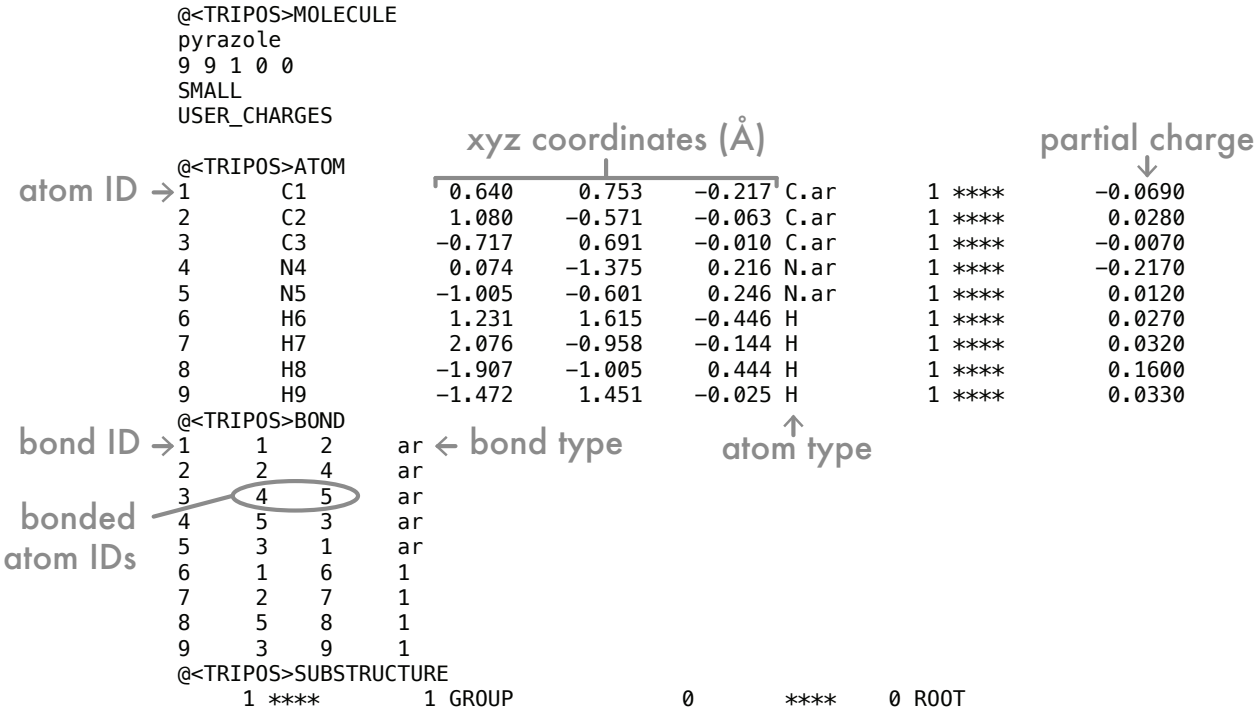
There exist a very large number of computational and mathematical techniques for optimising molecular geometries and calculating partial charges, and some of these are described in more detail in Section 4.2.1.1.[262,265] However for the purposes of this initial implementation, and for testing the hypothesis that bioisosteres can be retrieved and discovered through the searching of VEHICLe by shape and ESP similarity, geometry optimisations and partial charge calculations were carried out using the semi-empirical extended tight-binding (xTB) model described by Grimme *et al.* in 2017.[261] This model is designed to balance computational speed with physical accuracy, employing a semi-empirical approach that approximates the Hamiltonian using precomputed or experimentally derived parameters. It is based on the linear combination of atomic orbitals (LCAO) method, where atomic orbitals serve as the basis functions for constructing molecular wavefunctions. xTB is easily accessed using autodE as a wrapper, and is available to download as a Python package from common package managers, so installation is straightforward for users. Within this method, geometries are

optimised using the ANCOpt method at the GFN2-xTB level of theory, and partial charges are extracted from a single-point calculation using Mulliken population analysis.[261]

Geometry optimisations and partial charge calculations can be carried out either in the gas-phase (*in vacuo*) or in the condensed phase using a solvent model such as a polarisable continuum, where the solvent is treated as a continuous dielectric medium with the dielectric constant ϵ adjusted to match that of the desired solvent.[267] The inclusion of solvation can be important for molecules with many degrees of conformational freedom or those that are highly polar, as the solvent environment can perturb the electronic distribution or favour certain conformers. However, as VEHICLe consists of small and rigid aromatic heterocycles, the impact of solvation on either the optimised geometries or the calculated partial charges is likely to be negligible.[267] Accordingly, unless otherwise stated, all geometry optimisations and partial charge charge calculations throughout this thesis are carried out in the gas-phase for computational efficiency.[267,268]

2.1.3.3 Mol2 Output File

Once the geometry has been optimised and partial charges calculated, this information must be provided to the ShaEP algorithm alongside information about the connectivity of the atoms and the nature of these bonds. There are numerous file formats currently used in bio- and cheminformatics for representing molecules, and they each have their respective strengths and weaknesses, however the TRIPOS `.mol2` file is the only molecular file format accepted as an input by the ShaEP algorithm that includes information about the atomic partial charges within the molecule.[269] An annotated example of this format is shown in Figure 2.5.




```
@<TRIPOS>MOLECULE
pyrazole
9 9 1 0 0
SMALL
USER_CHARGES

@<TRIPOS>ATOM
1       C1          0.640     0.753    -0.217 C.ar      1 ****       -0.0690
2       C2          1.080    -0.571    -0.063 C.ar      1 ****        0.0280
3       C3         -0.717     0.691    -0.010 C.ar      1 ****       -0.0070
4       N4          0.074    -1.375     0.216 N.ar      1 ****       -0.2170
5       N5         -1.005    -0.601     0.246 N.ar      1 ****        0.0120
6       H6          1.231     1.615    -0.446 H         1 ****        0.0270
7       H7          2.076    -0.958    -0.144 H         1 ****        0.0320
8       H8         -1.907    -1.005     0.444 H         1 ****        0.1600
9       H9         -1.472     1.451    -0.025 H         1 ****        0.0330
@<TRIPOS>BOND
1     1    2     ar
2     2    4     ar
3     4    5     ar
4     5    3     ar
5     3    1     ar
6     1    6     1
7     2    7     1
8     5    8     1
9     3    9     1
@<TRIPOS>SUBSTRUCTURE
     1 ****        1 GROUP            0       ****    0 ROOT
```

Figure 2.5. An example of a `.mol2` file for pyrazole, annotated with its key features.

The `@<TRIPOS>MOLECULE` block encodes information about the molecule, including the name given to it by the user, the number of atoms, the number of bonds, the type of molecule (small molecule, protein, nucleic acid etc.), and the type of charges contained within the file. In the cases documented here these will all be `USER_CHARGES`.

The `@<TRIPOS>ATOM` block encodes all the necessary information about each atom in the molecule, including a unique identifier, the 3D coordinates (after optimisation), the type of atom (aromatic, H, $sp/sp^2/sp^3$), and the partial charges. The `@<TRIPOS>BOND` block then encodes all the information about the bonds in the molecule, including which atoms are bonded to each other, and the nature of that bond (aromatic, single, double etc.). The `@<TRIPOS>SUBSTRUCTURE` block is included for completeness, but has no relevance for the operations in this implementation.

These files are generated by HCIE for each query molecule, and are exported with the results for inspection by the user. For each of the 24 867 VEHICLe heterocycles, these were optimised and their geometries calculated, and then all this information collated into a single, 22 MB, multi-molecule `.mol2` file that is distributed with the package and used as the searchable library. The `.mol2` file for the query molecule is generated and exported on-the-fly.

2.1.4 Searching the Database

As aforementioned, the ShaEP algorithm by Vainio *et al.* was employed in this implementation as the searching and scoring algorithm.[216] This section outlines how this algorithm works to align and score the probe molecules in the VEHICLe database to the user-specified query molecule, and return the alignment of highest similarity to the user. A significant disadvantage in the use of ShaEP for this application is that its methodology is very complicated and so certain details, especially a detailed treatment of the mathematics involved in finding the optimal alignment, are alluded to but not described in any detail herein. Readers requiring a more detailed description of the mathematics are referred to the original publication in reference [216], where a more thorough treatment is given.

The ShaEP algorithm uses both a volumetric and an interaction field-based approach to identifying similarity. Both the query molecule and the library of probe molecules (in this case the VEHICLe database) are provided as `.mol2` files with pre-optimised geometries and pre-calculated partial charges. A field-graph (*vide infra*) is constructed according to predefined rules around each molecule, and the nodes of this are labelled with descriptors of the ESP and shape of the molecule. The maximal common subgraph (MCS) between the query field-graph and the probe is found, and rigid-body transformations that align these

subgraphs are found. These transformations are then clustered to eliminate any redundant transformations, before each of the remaining transformations is applied to the probe and optimised to maximise ESP and shape similarity. After optimisation the highest scoring alignment is then returned to the user. This process is repeated for each probe in the library. The highest score for each probe molecule in the library is returned to the user, and the alignments for the top 10 molecules are returned in an `sdf` file that can easily be viewed in molecular visualisation software. This process is illustrated in Figure 2.6.

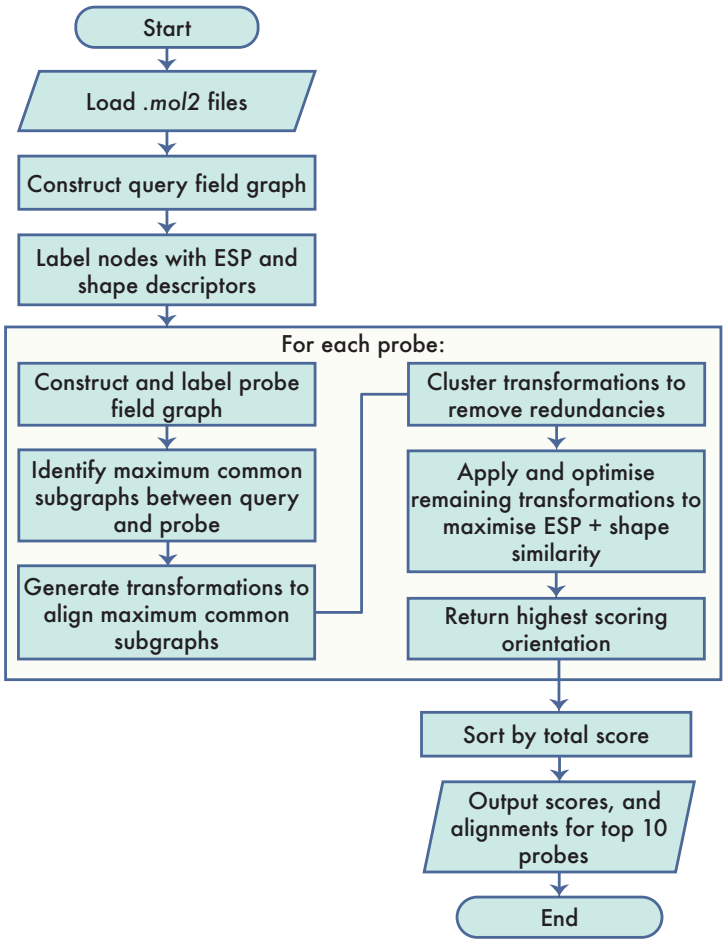


Figure 2.6. A flowchart demonstrating the ShaEP algorithm

2.1.4.1 Field-Graph Construction

Once the `mol2` file for the query has been read-in, a field graph is constructed around the molecule. A field-graph is a graph-based representation of a molecule's electrostatic potential (ESP) and shape, where nodes (vertices) are positioned in spatially defined regions around the molecule based on atomic positions, and edges describe relationships between these points. This graph is designed to provide an approximate representation of how the molecule interacts electrostatically with its environment, rather than just capturing atomic charges. By evaluating the ESP at these strategically placed points around the molecule, the cumulative influence of multiple atomic charges and the distance-dependent decay of electrostatic interactions can be effectively captured.

A field-graph is constructed around a molecule according to the rules given below. In the ShaEP implementation used here, field-graph nodes are positioned at a distance of $\sigma + h$ unless otherwise indicated, where σ is the van der Waals radius of the origin atom, and h is a user-adjustable parameter that is set to 0.2 Å.

1. Atoms bonded to exactly one non-hydrogen neighbour (for example a hydrogen or a monovalent halide) have a field-graph node $\sigma + h$ from the original atom, in the direction opposite to the non-hydrogen neighbour.

2. For non-hydrogen atoms, nodes are added in place of absent bonded atoms that would be expected for the hybridisation state of the atom in question:

    a) sp^3 hybridised atoms should have four neighbouring atoms in a tetrahedral geometry. Field-graph nodes are added in place of absent neighbouring atoms.

b) sp^2 hybridised atoms should have three neighbouring atoms in a trigonal planar geometry. Nodes are placed in the position of missing neighbours.

c) sp hybridised atoms should have two neighbours in a linear geometry. Place nodes to account for any absent neighbours.

3. Planar rings with more than four atoms have nodes placed above and below the plane of the ring, 1.6 Å from the centre on a line that is parallel to the normal of the ring plane and passes through the ring centroid.

An illustration of the construction of a field-graph for a triazine example molecule is shown in Panel A of Figure 2.7.

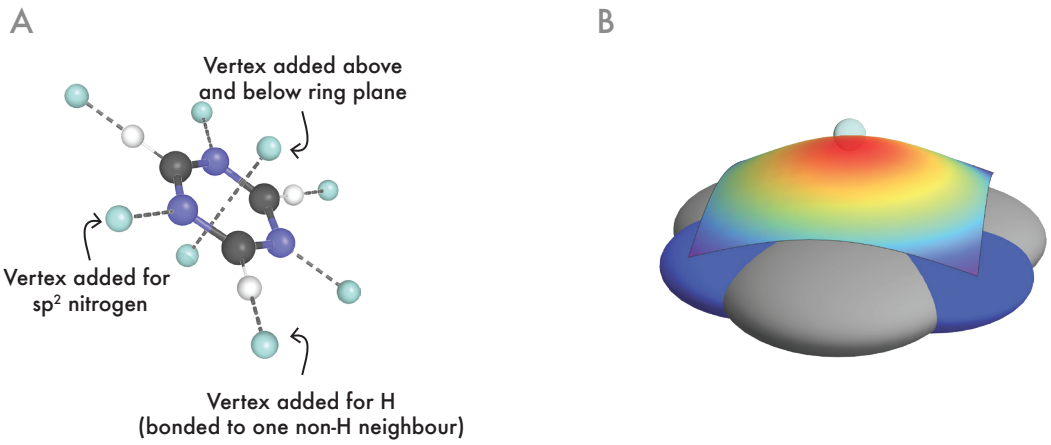


Figure 2.7. An example of a field-graph, and an illustration of the molecular shape-density surface for a single field-graph vertex.

Once the field-graph has been constructed, any nodes that lie within the van der Waals radius of any atoms are removed, and any nodes that are closer than 1.0 Å to each other are clustered together. Each remaining node is then labelled with a measure of the value of the ESP and a description of the local shape of the molecule at the node. The ESP value at the node is calculated using Coulomb's law

$$\phi_{\mathrm{ESP}} = \frac{1}{4\pi\epsilon_0\epsilon_r}\sum_i \frac{q_i}{d_i} \tag{2.1}$$

where q_i and d_i are the charge on atom i and Euclidean distance between atom i and the field-graph node respectively, ϵ_r is the relative static permittivity of the medium (which defaults in this case to 1), and ϵ_0 is the absolute dielectric permittivity of a vacuum. The summation runs over all atoms in the molecule, and gives a scalar value of the ESP at the field-graph node position.

To give a measure of the molecular shape at the node, a local shape descriptor is calculated at each node. This involves placing 3D Gaussian functions at the atomic coordinates of each atom

$$\rho_i(\vec{\boldsymbol{r}}) = 2\sqrt{2}\exp(-\alpha_i(\vec{\boldsymbol{r}} - \vec{\boldsymbol{R_i}})^2) \tag{2.2}$$

where $\vec{\boldsymbol{R_i}}$ are the coordinates of atom i. The authors take the decay factor α_i directly from Grant and Pickup (1995):

$$\alpha_i = \pi\left(\frac{6\sqrt{2}}{4\pi\sigma_i^3}\right)^{\frac{2}{3}} \tag{2.3}$$

where σ_i is the van der Waals radius of atom i.[270] A shape density surface is then created at each node by summing together each atom density function. This is illustrated for a single field-graph node in Panel B of Figure 2.7.

To describe the molecular shape at the node, the normal vector $\vec{\boldsymbol{n}}$ to the tangential plane of the shape density surface at the node is calculated, and the signed distance of each atom i from the plane is calculated using $\vec{\boldsymbol{n}} \cdot \vec{\boldsymbol{R}}_i - \vec{\boldsymbol{n}} \cdot \vec{\boldsymbol{r}}$. These distances to each atom from a node are then binned according to their sign and magnitude into a histogram vector (this is simply a vector representation of a histogram, where each row corresponds to a bin, and the value within it the number of atoms whose distance falls in that bin). This vector is then normalised to unit length. These are calculated for each node, thus each node on the field-graph is labelled with a scalar ESP value, and a unit-length histogram vector.

The advantage of using a histogram vector is that it is able to characterise the curvature of the molecular volume at each node, and therefore can give an indication of the smoothness of the molecular geometry. For example a planar molecule (like an aromatic ring) would have flat, symmetric histogram vectors, whereas a sterically demanding molecule with a high degree of three-dimensionality would have broader, more multimodal histogram vectors.

In the final field-graph, each node is labelled as above, and each edge is labelled with the Euclidean distance between the nodes it connects. The field-graph is therefore a complete graph, with an edge connecting each pair of nodes.

2.1.4.2 Subgraph Matching and Transformations

Once the completely labelled field-graph is constructed for both the probe and the query molecule, the maximal subgraph isomorphism between the query graph and the probe graph is located. In this context the maximal subgraph isomorphism (or maximum common subgraph; MCS) refers to the largest possible subgraph that exists in both the query and the probe's field-graphs, preserving the node connectivity and the labels. This concept is illus-

trated simply in Figure 2.8, where the outer graphs are complete graphs, with the colour of the nodes representing the labels. The graph in the centre is the MCS of both graphs, as all its node-colours appear with the same connectivity in both the outermost left and outermost right graph. A description of the precise algorithm by which these are identified is beyond the scope of this thesis, and the interested reader is referred to the original publication of Krissinel *et al.* for a thorough description.[271]

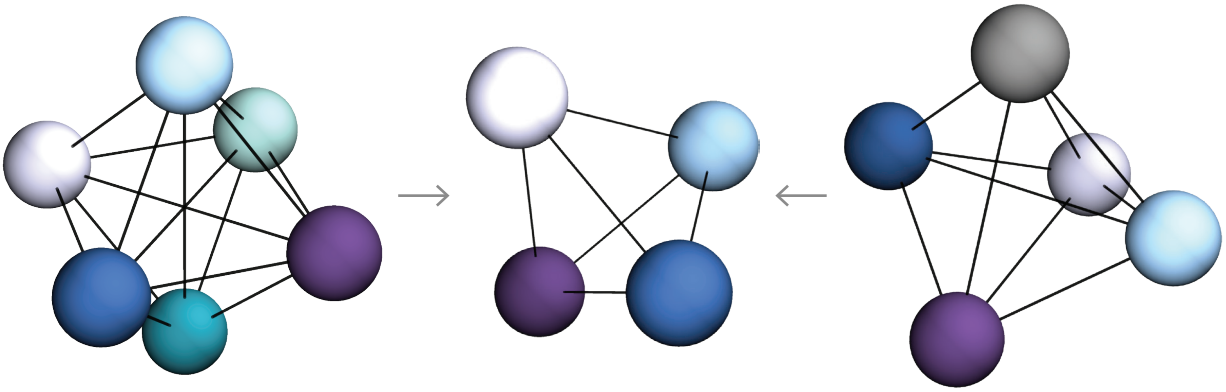


Figure 2.8. An example of a maximal subgraph. The graph in the centre is the maximal subgraph common to both complete graphs to the left and the right. As the nodes of each graph in ShaEP are labelled with ESP and shape descriptors, so the MCS captures the largest region of topological and electrostatic similarity between the molecules, and is used to guide alignment and scoring.

When identifying MCSes, the values of the labels at the nodes and vertices need to be compared. In the ShaEP algorithm used in this implementation, two edges or nodes are considered to be matching if:

- Their ESPs differ by at most 0.5 V and,
- The dot product of their shape histogram vectors is ≥ 0.866.
- For edges, their difference in distance must be ≤ 1 Å.

When the MCSes for the query and the probe are identified, rigid-body transformations that map those of the probe onto the target are generated, and those transformations that are very similar are clustered together and replaced with the average transformation for that cluster, calculated using dual quaternion algebra[c] using the method described by Kavan *et al.*[272]

Each of the transformations remaining after clustering is then applied to the probe molecule, and the resulting alignment scored for ESP and volumetric overlap similarity. For precise details of the scoring mechanism readers are referred to the original publication in reference [216], but the average mean of the ESP and volume overlap is calculated. Each alignment is then optimised to maximise this combined score, and the resulting alignment with the highest maximised score is returned to the user, with the final score reported as the Hodgkin similarity index, normalised to sit in the interval $[0, 1]$.[273]

This process is repeated for each probe molecule in the VEHICLe library, and the highest score for each probe is returned, along with an `sdf` file containing the alignments of the top 20 highest-scoring probes.

2.1.5 Visualisation

As the key exit-vectors in the highest similarity alignments are not indicated in the results output, it is necessary for the users to be able to visualise these alignments to decide on the key exit-vectors and design compounds based on the results. Distributed with the HCIE

[c]Quaternions are extensions of complex numbers into 4 dimensions and are an efficient mechanism for representing rotations. Dual quaternions are an extension of this number system to include both rotation and translation (and thus encapsulate stably and succinctly any rigid body transformation). Their algebra is complicated, and well outside the scope of this thesis.

package is a simple PyMol script enabling the easy viewing of these aligned structures. An example of this is shown in Panel **A** of Figure 2.9.

The codebase was written as a Python package, freely distributed from GitHub, and easy to install. The package was written to be object-oriented, and easy to use for those with little-to-no Python experience. For example a query molecule can be instantiated, its geometry optimised, partial charges calculated, and a ShaEP search of the query through the VEHICLe database in just two lines of Python code, as demonstrated below.

```
from hcie import Molecule

triazine = Molecule('c1ncncn1', name='triazine')
triazine.shaep()
```

2.2 Results and Discussion

2.2.1 Triazine

The results of a HCIE search for triazine, as illustrated in the code snippet above, are shown in Figure 2.9. This search took 7 minutes and 43 seconds (an average of 18.65 ms per structure).[d] Panel **A** shows the structural alignments returned by the ShaEP algorithm in the `sdf` file, and visualised using the PyMol script distributed with HCIE. Panel **B** shows a skeletal representation of these structures for clarity, with the RegIDs of each molecule in the VEHICLe database and their total score displayed beneath.

Reassuringly the top returned molecule from the VEHICLe database was the triazine query itself with a score of 1.0, implying perfect similarity with the query molecule. It is interesting

[d] Details of the computational configuration are provided in Section 8.1.

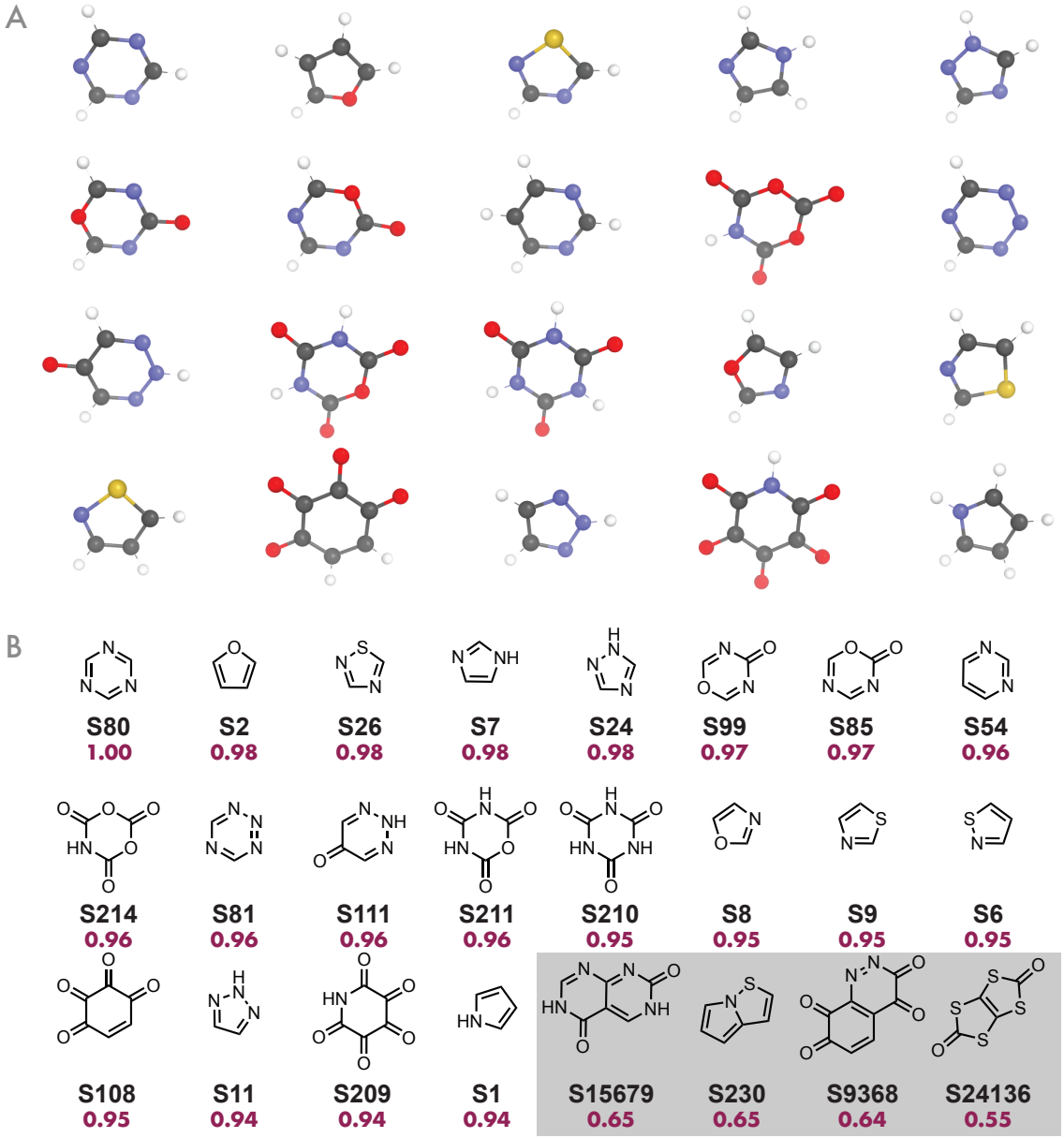


Figure 2.9. The results of a HCIE search for triazine. Panel **A** shows the aligned output structures, as given in the `sdf` file. Panel **B** shows the skeletal formulae for these structures, labelled by their RegID in the VEHICLe database and their total ShaEP score. The compound order is preserved in both panels. Also shown highlighted in grey are four of the lowest scoring molecules returned in the search.

to note that the scores are universally very high across these top 20 molecules, with all the scores in the range [0.94, 0.98]. It is also intriguing to note that nine of the top 20 proposed bioisosteres of the 6-membered query triazine are five-membered rings (all with very high

similarity scores). Although the bioisosteric substitution of six-membered rings for five-membered rings is commonplace in medicinal chemistry, the difference in size between these two classes of ring system is large enough that a greater difference in similarity score between these would be expected.[81]

It is also interesting to note that whereas the query molecule has a C_3 proper rotation axis, only **S210** in the results shares this symmetry operation (although **S214** and **S211** almost share this operation). The top proposed bioisostere **S2** does not share this symmetry operation, thus its high similarity score is surprising, as the electronic distribution of the ideal bioisostere should share as close as possible the symmetry of the ESP of the query molecule.

Also worthy of note in these results is the inclusion of several molecules that are likely to be too unstable to exist more than fleetingly, or be medicinally useful. **S209** and **S108** stand out amongst these, however it is likely that **S214**, **S111**, **S211**, and **S210** are also too reactive or unstable to exist more than fleetingly. Whilst the inclusion of such molecules in VEHICLe is unavoidable, and indeed the inclusion of previously unsynthesised molecules in the searchable library is crucial to broaden the range of bioisosteric chemical space, these molecules appearing within the results makes prospective compound design difficult without further filtering, and potentially obscures more realistic bioisosteric candidates by pushing them further down the ranked results.

Displayed in Figure 2.9 panel B alongside the top ranked molecules are four examples of heterocycles that were scored poorly against the triazine query, highlighted in grey. Reassuringly these all appear by inspection to be significantly different to the triazine query; all are bicyclic molecules and are therefore more sterically demanding than the monocyclic

query, and all are likely to have highly asymmetric electronic distributions, in particular **S230** and **S9368**. That these molecules score poorly against the triazine query suggests that the ShaEP scoring algorithm is effectively capturing the key features of molecular similarity, deprioritising heterocycles with steric and electronic profiles that deviate markedly from the query structure.

2.2.2 SARS-CoV-2 Main Protease Inhibitors

In order to benchmark the performance of this implementation of HCIE in retrieving known active compounds and proposing novel bioisosteric pairings, an appropriate dataset is required. This dataset must possess the following properties:

1. Have a sufficiently large number of matched molecular pairs (MMPs) that any correlations or results drawn from it are statistically reliable.

2. Each of the MMPs in the dataset should be simple aromatic ring substitutions on a common core, with each of the aromatic rings in question being a member of the VEHICLe database.

3. Each of these MMPs should be appropriately biologically characterised in a reliable and reproducible assay that permits comparison within the dataset.

4. Ideally there should be significant structural information about the binding modes and poses of these ligands, such that they can be scored in these binding alignments, and these compared to the 'pose-blind' alignments generated by the ShaEP alignment.

Identifying a dataset in the literature that met all of these criteria proved to be challenging. Eventually the series of inhibitors of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) main protease (MPro) developed and released by the COVID Moonshot project were selected as the benchmarking dataset.[274]

SARS-CoV-2 is the causative pathogen of the 2020 COVID-19 pandemic, which as of February 2025 is estimated to have infected over 770 million people worldwide, causing or contributing to 7 million deaths globally.[275] Identifying small-molecule anti-infectives for this highly contagious global pathogen thus became a challenge of extreme importance for medicinal chemistry.[274] When designing anti-infective therapeutics the best targets are those that have no host analogues and are essential to the pathogen. The SARS-CoV-2 MPro is an essential coronavirus enzyme that cleaves viral polyproteins into small non-structural proteins, several of which form the viral replication-transcription complex which synthesises new viral RNAs.[276,277] Structural characterisation of the SARS-CoV-2 MPro reveals it to be a 33.8 kDa homodimer of two 306-residue protomers, with a two-fold rotational axis of symmetry relating two identical small-molecule binding regions in the active site, as shown in Panel **A** of Figure 2.11.[278] Protease activity is facilitated by a cysteine-histidine catalytic dyad cleaving peptides after a glutamine residue, a sequence-specificity for which there is no known human analogue.[276]

The COVID Moonshot project was an open-source collaboration between academic and industrial scientists that aimed to crowdsource the design of small-molecule inhibitors of the SARS-CoV-2 MPro, and was co-founded in 2020 at the Centre for Medicines Discovery in Oxford. It aimed to synthesise, and structurally and biologically characterise these using computational screens, automated synthesis platforms, and robotic assay and X-ray crystallography workflows.[274] The objective was to significantly decrease the time taken to develop

a potent, drug-like inhibitor by publicly releasing all of the data in real time, and inviting rational-design suggestions from the medicinal chemistry community based on this data. The workflow, along with the lead candidate to emerge from this pipeline, is shown in Figure 2.10 (this figure is reproduced in its entirety from reference [274], with permission from AAAS).

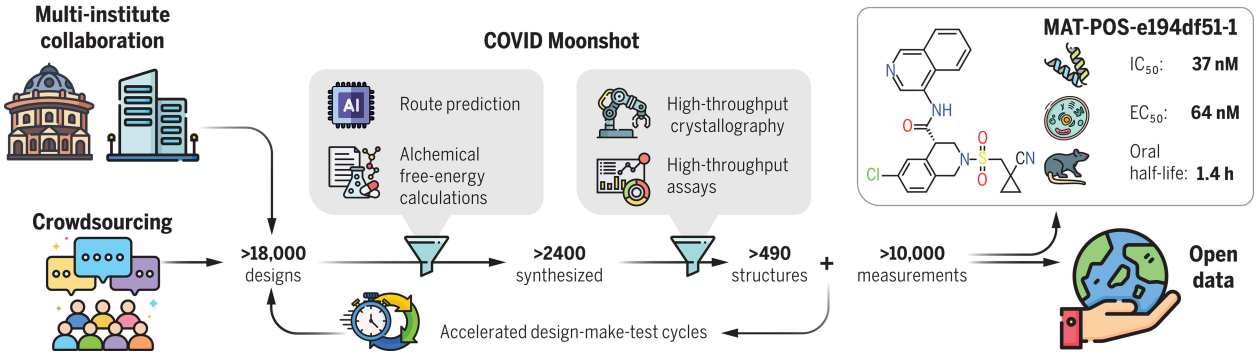


Figure 2.10. The COVID Moonshot pipeline. Figure is taken from Boby *et al.* (2023).[274] Reprinted with permission from AAAS.

The authors were able to synthesise over 2400 compounds, resulting in more than 490 small-molecule X-ray crystal structures and over 10 000 biophysical and biochemical measurements. All of these data were released into the public domain, and thus this dataset was initially mined for matched molecular series (MMS) that meet the specified criteria.

Within this dataset were a series of 372 inhibitors based on a chlorobenzylacetamide core scaffold, of which 35 of these molecules represented an MMS of only unsubstituted aromatic heterocycles that were present in VEHICLe and were annotated with IC_{50} data. The ligands defining this series are shown in Panel **B** of Figure 2.11. Of these, 13 had crystal structures which showed that the heterocyclic portions of each ligand lie in the same position and plane within the binding site. An example of this binding mode is shown for the isoquinoline derivative **2.1p** in Panel **A** of Figure 2.11.

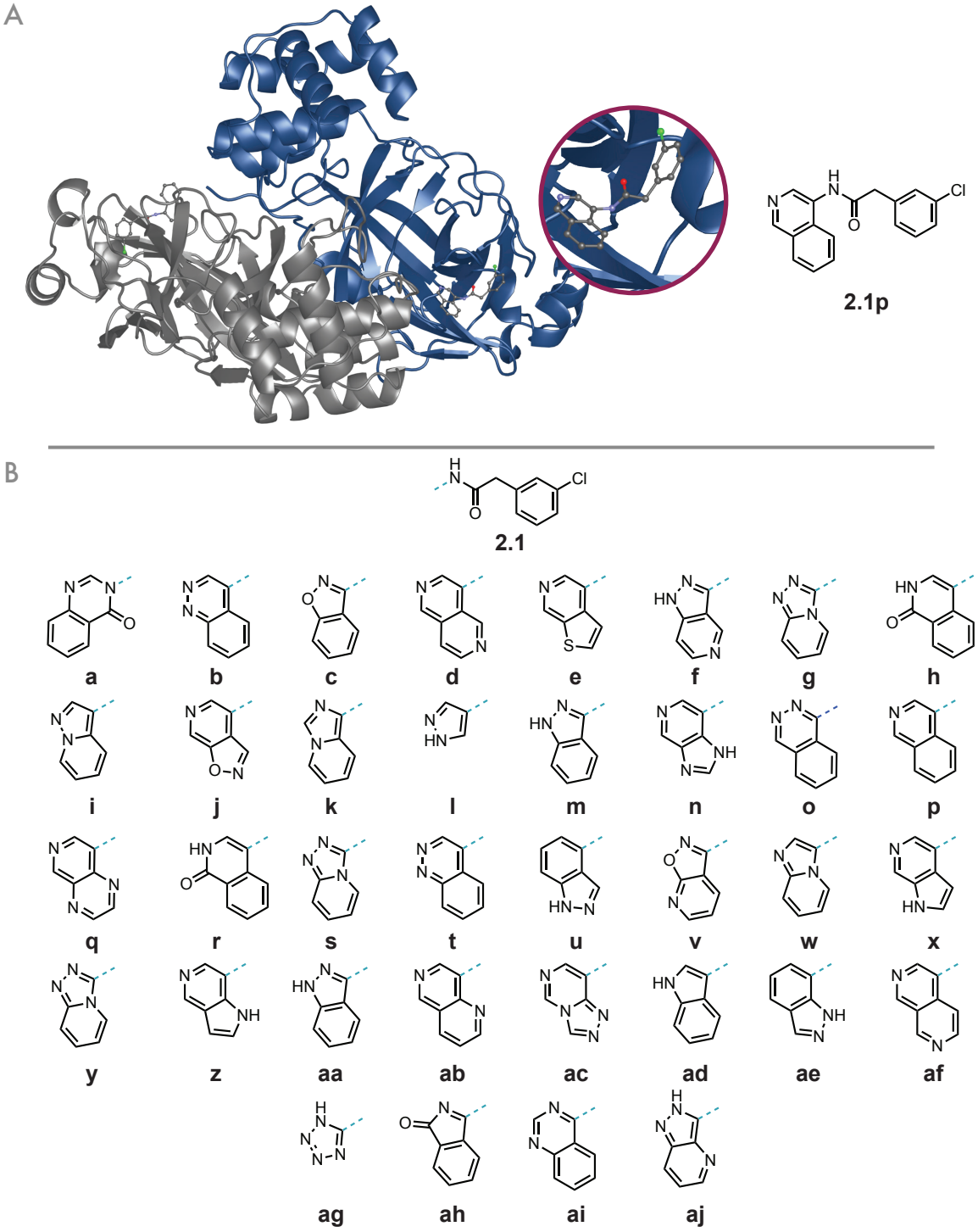


Figure 2.11. The chlorobenzylacetamide series of COVID MPro inhibitors discovered by the COVID Moonshot project. Panel **A** shows the crystal structure of an isoquinoline-derived inhibitor (Moonshot ID: ADA-UCB-6c2cb422-1) in this series bound to the MPro homodimer (PDB: 7GAV). Panel **B** shows the heterocyclic analogues from the series used in this analysis.

2.2.2.1 Bioactivity Benchmarking

To investigate the validity of the HCIE and ShaEP-based search approach, and the key assumption of bioisosterism presented in Figure 2.1, attempts were made to correlate the bioactivities of the ligand series depicted in Figure 2.11B to their similarity scores as determined by ShaEP. The effect of the method by which the partial charges were determined was also investigated.

The bioactivities for the ligands in the chlorobenzylacetamide series were given as IC_{50} values determined by fluorescence inhibition assay. The authors incubated each inhibitor in the series with isolated SARS-CoV-2 MPro for 15 minutes at room temperature, before a substrate peptide with a fluorophore (5-FAM) and a fluorescence quencher (Dabcyl) at opposing ends were added and incubated for a further half hour before a measurement of fluorescence is taken. In the native peptide the quencher is sufficiently close to the fluorophore that no fluorescence is observed, however if the peptide is cleaved by MPro then the quencher is no longer able to suppress the fluorescence and a quantifiable luminosity is observed. By repeating this assay at a series of inhibitor concentrations, the authors constructed dose-response curves and obtain IC_{50} values for each inhibitor. These were used as the measure of bioactivity in the below analysis.

The ligands were prepared by aligning each of the chlorobenzylacetamide cores with those for which X-ray crystal structures existed. Each of the aromatic rings was then aligned artificially to lie in the same plane and direction as the heterocyclic portions of the X-ray characterised molecules. These alignments were carried out using a flexible body alignment algorithm implemented in the Schrödinger® software package. These geometries were exported to a `mol2` file and partial charges calculated and included. This multi-molecule `mol2` file was then used as the searchable library for a HCIE search, with ShaEP set to score the

provided alignments, but not to alter their coordinates. Thienopyridine derivate **2.1e**, the most potent in the series with an $IC_{50} = 0.73$ µM, was used as the query molecule and all of the other ligands scored against it for similarity in the manner described above.

These similarities were then plotted against the difference in pIC_{50} from the most potent ligand (**2.1e**). Several of the ligands in the series had IC_{50} values > 99.5 µM, which is the value at which the fluorescence assay saturates, and as such these were categorised as 'inactive'.[e] Ligands with an IC_{50} lower than this were classed as 'active'. These correlations are shown in Figure 2.12, where the abscissa measures the similarity score compared to **2.1e**, as calculated by ShaEP, and the ordinate measures the difference in pIC_{50} from **2.1e** for each ligand in the series. This analysis is based on the hypothesis that compounds in the series with a high level of similarity both volumetrically and electronically should have similar pIC_{50} values, and therefore be bioisosteres in terms of potency.

To observe the effect of the method of partial charge assignment on these correlations, partial charges were calculated for the ligands using three different electronic structure methods representing low, medium, and high levels of theory. Charges calculated by the parameterised method introduced by Gasteiger and Marsili in 1980 and implemented efficiently in RDKit represented the low-level approach, the semi-empirical GFN2-xTB charges (partitioned using Mulliken population analysis) discussed previously represent a mid-tier level of theory, and DFT charges calculated in the ORCA electronic structure package using D3BJ-PBE0/def2-TZVP functionals represented the high-level, computationally costly level of theory. A more detailed discussion of the precise details of these various levels of theory is given in Section 4.2.1.4. Three separate `mol2` library files were prepared containing the charges and aligned

[e]No solubility issues were observed under the assay conditions, so reported inactivity is unlikely to be due to lack of solubility.

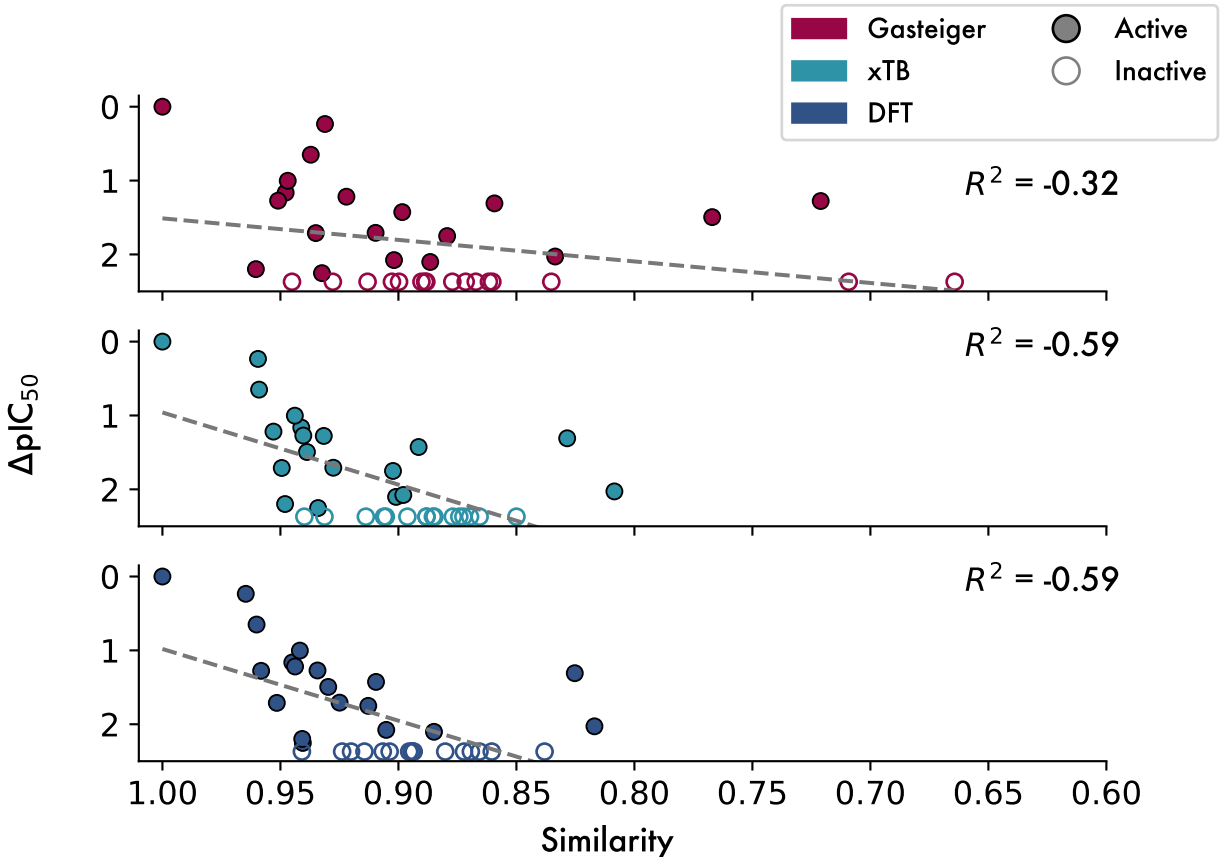


Figure 2.12. The difference in pIC_{50} against the ShaEP similarity for the heteroaromatic ligands in the chlorobenzylacetamide series of MPro inhibitors for different partial charge calculation methods. The dashed lines in grey represent lines of best fit, and the R^2 values are Pearson correlation coefficient values.

geometries of all 35 ligands, and these were scored against the most potent ligand in the manner described above. The results of these analyses are also shown in Figure 2.12.

All of the similarity scores are reasonably high, falling in the range [0.65, 1.00] for the Gasteiger set, and in the narrower [0.80, 1.00] range for the xTB and DFT sets. This is partly attributable to the inclusion of the common chlorobenzylacetamide core when scoring the ligands. However, as was shown in Section 2.2.1 these scores tend to be high even for heterocyclic ligands without a common core. It is likely that the histogram vector method of characterising the molecular shape is better suited to larger molecules with a greater 3D

component, where significant differences in molecular shape are expected. For these ligand series, and indeed all the heterocycles in VEHICLe, the differences in the atomic positions are likely to be small (of the order of 1 Å for five-membered vs six-membered rings, and smaller than that for single heteroatom substitutions). The 1 Å bin sizes in the histogram vector assignment could therefore miss the subtleties of these small differences in atomic position, which may well capture effectively differences in shape for larger molecules, but is perhaps less effective at identifying these crucial smaller differences that are so important in this application.

It is clear from Figure 2.12 that there is some similarity score dependence on the electronic structure method, with the Gasteiger plot showing a lower correlation and different distribution of similarity scores to that of the xTB and DFT plots, which are more similar to each other. This is not entirely unexpected, as the parametrised Gasteiger calculations are conformationally independent, whereas the xTB and DFT calculations both take into account structural conformation. That xTB and DFT are similar is not surprising, given the significant amount of shared underlying theory between the two methods, and this similarity reassuringly supports the initial decision to use this method in the implementation, rather than the much slower DFT method.

There is some evidence of a linear relationship in Figure 2.12, especially in the xTB and DFT plots, suggesting that some correlation exists between the similarity score and the potency of the compounds. However, the handling of inactive compounds is more scattered; the majority have similarity scores in the region 0.85 – 0.90, which is towards the lower end of the scores in the xTB and DFT distributions, however there are 5 inactive compounds in the xTB distribution with similarity scores > 0.90. In order of decreasing similarity these are **2.1b**, **2.1ae**, **2.1u**, **2.1ai**, and **2.1c**. Although without structural information it is difficult

to postulate a reason for these inactive outliers, it is possible that these ligands contain certain pharmacophoric features that are incompatible with the MPro binding pocket; for example **2.1b** and **2.1ai** both have an H-bond accepting nitrogen at the 4-position which could contribute to the reduction in binding affinity.

2.2.2.2 Enrichment Factors

To quantify HCIE's ability to differentiate between active and inactive compounds in this dataset, enrichment factors were calculated for each electronic structure method. Enrichment factors are a common metric for evaluating the potential of virtual ligand alignment and retrieval tools.[279] Calculation of these often involves taking a large database of non-bioactive ligands and seeding within it a small number of known active ligands. This database is scored and ranked with the tool being evaluated, and the ratio of the number of active ligands returned by the tool in the top fraction of results compared to what would be expected if the database was randomly sampled is calculated. More formally,

$$\mathrm{EF}_{\text{top n\%}} = \frac{\left(\dfrac{n_{\text{active}}}{n_{\text{total}}}\right)_{\text{top n\%}}}{\left(\dfrac{n_{\text{active}}}{n_{\text{total}}}\right)_{\text{database}}} \tag{2.4}$$

thus an enrichment factor of 1.0 implies that the screening methodology is no better than random sampling at retrieving active ligands, and a higher enrichment factor suggests that the probability of there being an active ligand in the top n% of the ranked database is greater than random. It is necessary to specify the top n% as, by definition, the enrichment factor

over the whole dataset is 1.0. These were calculated for the data in Figure 2.12, and the results are shown in Figure 2.13.

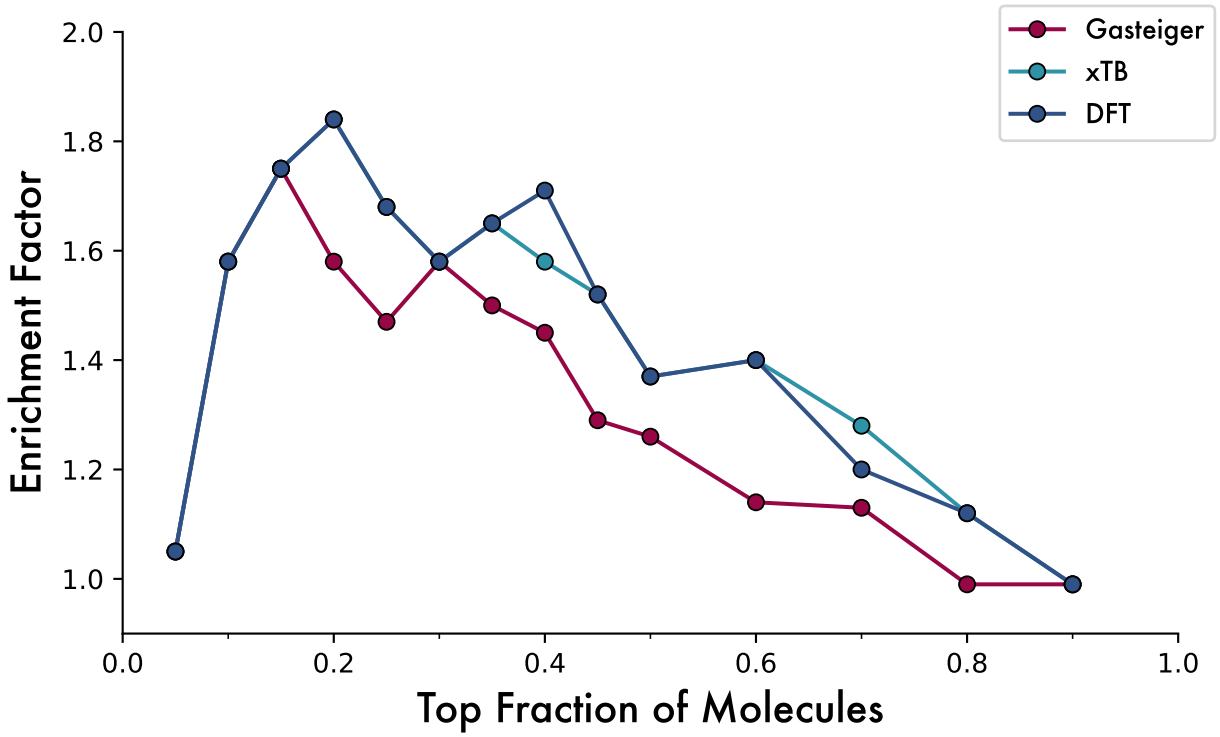


Figure 2.13. Enrichment factors for the chlorobenzylacetamide MPro inhibitor series, as determined by the ShaEP similarity scores for each of the three electronic structure methods.

As with the linear correlations, there is a greater agreement between the enrichment factors calculated by xTB and DFT than with those calculated using the Gasteiger charges. The $EF_{20\%}$ for those calculated using xTB is 1.84, suggesting that the ShaEP scoring method is 1.84 times more likely to return bioactive ligands when compared to random chance. These enrichments are modest, but they must be viewed in the context of an already highly-enriched dataset (57% of ligands in this set are classed as active), which will reduce the magnitude of the EFs compared to a less-enriched dataset. Therefore, given the already high likelihood of there being a large number of active ligands in a random sample, an enrichment factor of 1.8 represents a significant improvement on random chance. When selecting or prioritising

compounds for synthesis and biological evaluation, focussing efforts towards those that are nearly twice as likely to be active represents a considerable improvement.

Although the above enrichments do represent an improvement over random chance, the seemingly poor differentiation between active and inactive compounds in Figure 2.12 was concerning. This suggested either that the method of similarity scoring used in this ShaEP-based implementation is insufficiently sensitive to differentiate between active and inactive ligands, or that there is no key volumetric or electrostatically defined pharmacophore that predicts well the ability of a compound to bind to the MPro active site. It is likely that a combination of the two is responsible for this. It has been mentioned earlier that the universally high scores for comparisons within VEHICLe, even when comparing volumetrically different molecules (for example furan to triazine), suggests that the histogram vector-based method of describing molecular shape is too coarse-grained to identify the small changes in atomic position that differentiate the various shapes of VEHICLe heterocycles.

ShaEP's algorithm is principally designed for the rigid-body alignment and similarity scoring of larger ligands with greater degrees of conformational flexibility than the small, unsubstituted heterocycles in VEHICLe. This is illustrated by datasets Vainio *et al.* chose to benchmark on, which all have large sp^3 proportions and flexible sidechains. The presence of some degree of linearity in the IC_{50} correlations shown in Figure 2.12 is encouraging, suggesting that there is merit in the hypothesis proposed in Figure 2.1, namely that similarities in shape and electrostatics can lead to similarities in biological activity, however these initial investigations demonstrate that a higher resolution approach is required to more precisely quantify shape similarity between the VEHICLe heterocycles.

There is also a growing body of published literature suggesting that *in silico* virtual screening campaigns should treat SARS-CoV-2 inhibitors with caution.[280–282] Molecular dynamics studies by Bzówka *et al.* revealed a high degree of flexibility in the ligand binding site of the MPro, suggesting that multiple binding site conformations exist.[281] This is likely to confound virtual screening efforts as there will not be one single ligand shape that defines well the binding site, and for an analysis such as this one it could be that there exist even within this chlorobenzylacetamide series several 'correct' binding poses. This will skew results both towards false negatives, where ligands of a different shape (and therefore low similarity) to the query ligand are actually strong binders to the target but in a different conformation, and false positives where inactive compounds score highly on similarity and are therefore predicted as active. These findings were corroborated by Macip *et al.*, who used literature MPro inhibitors, including those from the COVID Moonshot project, to determine whether a variety of docking scores and methods were able to distinguish between active and inactive inhibitors.[280] The authors found that the docking methods and scores were in general no better than random at differentiating between active and inactive compounds, reinforcing the effect observed here. This suggestion is strengthened by the work of Zev *et al.* who tested six of the most popular docking programmes against 193 MPro ligand binding poses determined by X-ray crystallography, finding that only 26% of these poses were predicted correctly.[282]

Taken together with the results described in Section 2.2.2.1, these suggest that the MPro dataset, although large, might not be the most appropriate set for benchmarking HCIE. None of the above sources suggest that there is no role for virtual screening methodologies in MPro inhibitor discovery, but each advise caution in interpreting the results of virtual screens and docking studies on this flexible target. In light of the factors discussed here, further

work on MPro inhibitors, including any experimental validation of these HCIE results, was discontinued.

2.2.3 Methodology Limitations

Whilst the ShaEP methodology provides a convenient and well-tested means of exploring the hypothesis of shape and electrostatic similarity contributions to bioisosterism, it does suffer from several limitations. As previously described, the characterisation of molecular shape as a histogram vector at each node of the field-graph assigns too high a similarity to aromatic heterocycles differing in ring size, thus too readily classifying five-membered rings as being highly similar to six-membered rings. More generally, the ShaEP codebase was written for larger and more structurally complex molecules, and the field-graph alignment and scoring method is too complex for the task of aligning heterocycles with few conformational degrees of freedom and 3D character. Occam's razor states pithily that the simplest possible construction is usually the best, and the mathematically complex algorithm implemented in ShaEP is likely to deter users who do not understand it, limiting the breadth of appeal of HCIE as a tool.

This implementation also has no means by which users can specify points of attachment on the query molecule. In the context of drug discovery, heterocycles are invariably substituted at one or more positions.[283,284] The interactions of a drug molecule with its target are often largely determined by the geometrical positioning of key functionality from a core, as this enables the formation of specific non-covalent or covalent bonding interactions with the target residues. Thus when considering similar scaffolds for scaffold hopping or bioisosterically replacing a directly interacting heterocyclic group, the positioning of the exit-vector(s) about

the ring is a crucially important factor in determining whether a specific analogue will retain target-specific potency.

Furthermore it is possible that the best alignment of the proposed bioisosteres might not have an exit-vector in the correct orientation relative to the key vector in the query molecule. As the proposed bioisosteres are sorted by score, and the score is entirely a function of the relative alignment between query and probe molecules, this has the potential to skew the relative orders of the proposed bioisosteres. An example of a situation where this could be significant is illustrated in Figure 2.14. For a 2-substituted pyridine query molecule **2.2**, the best possible alignment of hypothetical heterocycle **2.3a** does not have an exit-vector in a suitable orientation for further functionalisation, and is therefore not a useful alignment, whereas the best possible alignment of **2.4** does have a vector in an appropriate orientation for further functionalisation, but this score is lower than that of **2.3a**'s optimal alignment so is ranked lower. However, the best alignment of **2.3** with a vector in an appropriate orientation (**2.3b**) might score lower than that of **2.4**. In this implementation **2.3a** would be ranked higher than that of **2.4** even though, when the appropriate alignment of exit-vectors is considered, **2.4** is a better-scoring isostere.

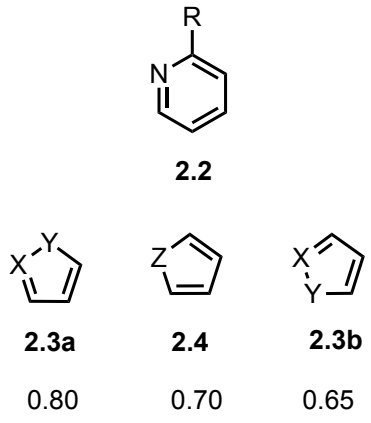


Figure 2.14. A situation where a non vector-based alignment leads to skewed results. The scores are arbitrary, and have been generated for illustrative purposes only.

The dependency on ShaEP for alignment and scoring also introduces an issue with regards to the availability of the source code. At the time of writing, the code is only available as a binary distribution, and therefore the code is not open-sourced. This leads to potential reproducibility issues, and moves the codebase away from the current best practices in research software engineering.[285,286] Osborne *et al.*'s 2014 publication *Ten Simple Rules for Effective Computational Research* establishes a set of simple and useful guidelines for developing and publishing scientific software which the authors suggest will improve the quality and reproducibility of scientific research conducted using software. Rule 9 of this is simply *share EVERYTHING* (sic), with the authors arguing that source code acts as a readable methodology of the approach undertaken, and thus it should be shared in full.[286] The use of a proprietary alignment and scoring algorithm also renders a solution to the exit-vector problem outlined above very difficult, as it is not possible to alter or amend the scoring algorithm without its source code. Therefore there exists a need to refactor the alignment and scoring methodology to remove this dependence on proprietary methodology.

2.3 Conclusions and Future Work

This chapter has described the development and initial testing of a first-generation implementation of the HCIE algorithm, making use of Pitt *et al.*'s VEHICLe database of aromatic heterocycles as a searchable library of potential bioisosteres.[256] A user-specified query molecule is searched against VEHICLe using the ShaEP alignment and scoring algorithm developed by Vainio *et al.*, which makes use of a field-graph based representation of the electrostatic surface potential and volumetric shape of the molecule, and uses an alignment algorithm based on the maximal common substructure of the field-graph nodes and edges in both the query and the probe molecule.[216] Query molecules and each of the VEHICLe

heterocycles had their geometries optimised and partial charges calculated using the xTB semi-empirical DFT model, and these were exported into a `mol2` file before searching. The search results are returned both as a list of VEHICLe RegIDs with their respective similarity scores, and an `sdf` file containing the coordinates of the top 10 highest-scoring ligands in their alignment of highest similarity.

The ShaEP-based HCIE method was benchmarked against a dataset of inhibitors of the SARS-CoV-2 MPro from the COVID Moonshot project, and the effect of the electronic structure method on the correlation between ShaEP similarity and bioactivity was also investigated. It was found that the correlations calculated using xTB charges and high-level DFT charges were in good agreement, but those calculated with the parametrised Gasteiger model gave weaker correlation. Although there was evidence of linearity in the relationship between bioactivity and similarity score, the wide range of scores exhibited by inactive compounds was a cause for concern. Similar analyses in the literature, combined with Bzówka *et al.*'s molecular mechanics investigation into the plasticity of the MPro binding site suggest that there exist multiple different chemotypes that bind in the same flexible pocket, and as such virtual screening and docking campaigns on these inhibitors might not yield wholly reliable results.[280–282] These factors, combined with the limitations on the ShaEP methodology for this application outlined in Section 2.2.3 above, lead to further work on MPro inhibitors being discontinued.

Alongside the limitations discussed in Section 2.2.3, another key limitation of this initial approach is that VEHICLe contains only unsubstituted heterocycles. Not only does this restrict the region of chemical space that the tool searches through, thus decreasing the likelihood of discovering novel, potent bioisosteres, but in doing so excludes a highly important class of molecule (functionalised aromatic heterocycles) from the searchable space.

Chapter 3 outlines the approach taken to include a range of substituents in the searchable library, thereby extending VEHICLe into the Mono-and-Bifunctionalised VEHICLe (MoBiVic) database. Chapter 4 describes the conception and implementation of a novel, vector-based alignment and scoring algorithm that addresses the concerns raised in Section 2.2.3.

Although ultimately the implementation described in this chapter was not the one used in the final package, these early findings informed significantly the subsequent work described in the remainder of this thesis. The insights gained into the constraints of the ShaEP methodology for the screening of small aromatic heterocycles shaped the design of the subsequent iteration of HCIE. The results demonstrated in Figure 2.12 also go some way to supporting a key foundational hypothesis of this work; namely that there is a correlation between bioactivity and similarity as determined by shape and electrostatics that can be exploited to discover novel heteroaromatic bioisosteres.

3 MoBiVic: The Expansion of the Virtual Library

The implementation described previously suffered from several shortcomings. This chapter aims to address the issue of the searchable heterocyclic library (VEHICLe) containing only unsubstituted heterocycles. The inclusion within VEHICLe of heterocycles that are likely to be too reactive or unstable to exist also skews the results generated, so a means of removing these unfeasible heterocycles is required. This chapter describes the expansion of the VEHICLe database to include monofunctionalised and bifunctionalised molecules, and also a means of filtering the generated dataset to remove unreasonable molecules.

3.1 Introduction

The use of VEHICLe in its original, unsubstituted form as a library significantly limits the region of chemical space that the tool can access, and by extension the usefulness of its results. Whilst VEHICLe represents a complete enumeration of the chemical space within the parameters set out by the original authors, aromatic heterocycles within approved drug molecules are often decorated with small substituents to improve physicochemical proper-

ties or potency, and thus including these molecules within the library would significantly expand its utility within drug discovery projects by providing access to a much larger region of chemical space with a broader variety of physicochemical properties.[58,79] As the electronic nature of the substituent can have profound effects on the overall electron distribution of ring atoms within aromatic rings, it is also highly likely that including substituted rings within the library would increase the chances of finding effective new bioisosteric pairings. Inclusion of electron withdrawing and donating substituents could alter the electron distribution within a ring such that a novel, substituted ring system could better mimic the electron distribution of a known, biologically active heterocycle. This is illustrated for the simple case of pyridine vs fluorobenzene in Figure 3.1, where the substitution of a fluorine in place of the aromatic nitrogen leads to a similar electrostatic surface potential, with the region of attractive electrostatic potential in fluorobenzene occupying a similar proportion of the ESP relative to the other atoms to that of the aromatic nitrogen in pyridine. Indeed this pairing has been widely used within the medicinal chemistry literature; searching SwissBioisostere at the time of writing for 2-pyridine reveals 591 published examples of this replacement, with 534 of these instances leading to improved or retained binding affinity.[193,219]

This concept of ring substituents with the correct electronic properties mimicking the electronic effects of heteroatoms within the ring is commonplace in synthetic chemistry, however often the ring substituents used to take advantage of this effect (for example a nitro group in place of a pyridine-like nitrogen) are themselves reactive and thus are flags for biological instability or lack of selectivity.[288] A substituted library would therefore have to be carefully compiled and curated to include substituents with a range of electronic properties, but that are not linked to undesirable biological mechanisms or known to be cross-reactive, such that

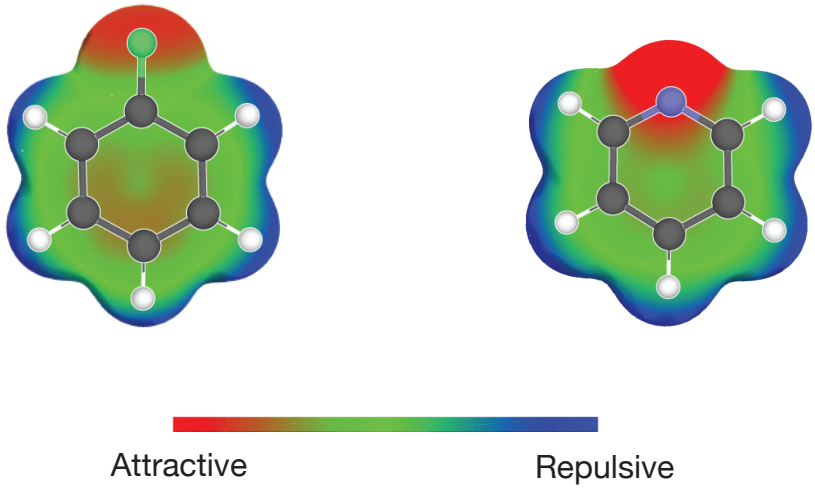


Figure 3.1. A visualisation of the electrostatic surface potentials of fluorobenzene (left) and pyridine (right). Electron density surfaces were calculated using DFT at the PBE0 level of theory with def2-TZVPP basis sets and a GD3BJ empirical dispersion correction.[287] Electrostatic surface potentials were generated on these electron density surfaces at 0.002 $\epsilon \cdot a_0^{-3}$, with regions of attractive potential visualised in red, and regions of repulsive potential in blue.

this effect can be as fully explored and exploited as possible, without introducing problematic functional groups to hit-like or lead-like molecules.

Whilst the examples outlined above are somewhat trivial, it was expected that these effects would still be relevant for more electrostatically complex ring systems (for example bicyclic rings with multiple donating and withdrawing ring atoms), and that for these molecules it would be far more difficult to predict mimetic arrangements of ring heteroatoms and substituents that lead to isosteric molecules. Thus a computational search for electrostatically similar molecules within an enumerated, substituted library of heterocyclic molecules has the potential to uncover new electrostatic bioisosteres of conventional or commonplace medicinally relevant heterocycles, thus further expanding the bioisosteric space available to medicinal chemists.

This chapter describes the development of MoBiVic (*the Mono and Bifunctionalised VEHICLe database*) library of heterocyclic molecules, based on mono- and bi-functionalising the VEHICLe database of Pitt *et al.*.[256] The filtering of this novel dataset using simple rules to exclude chemically unstable or medicinally unfeasible heterocycles is also described.

3.2 Methodology

The method used to select the substituents for and expand the VEHICLe database to broaden the region of chemical space that is accessible to the searching tool is described in this section, specifically:

- The criteria for selecting the substituents used to functionalise the library
- The computational workflow for generating the expanded database
- Filtering the generated library to remove highly reactive, or medicinally unfeasible molecules from the search results

The original VEHICLe database described by Pitt *et al.* in 2009 is unfunctionalised, meaning that the ring systems within it bear no substituents other than carbonyls.[256] As the unfunctionalised rings represent the core scaffold of any aromatic heterocycle containing molecule, it is a reasonable approximation to build a searchable library containing only these, thus leaving to the inspection of the end-user decisions on where to add further functionality. However, as demonstrated in Figure 3.1, the inclusion of substituents with differing electronic properties to that of the atom at the point of attachment can have a significant effect on the overall electron distribution of the aromatic ring, as recognised initially by Hammett in 1937.[289] This

perturbation of the overall electronics of the ring has the potential to alter the similarities of the returned ligands to a given query ligand if the proposed unfunctionalised bioisosteres are subsequently functionalised, thus altering their efficacy as bioisosteres. To correct for this, it was decided to expand the region of chemical space covered by the searchable library by methodically functionalising each molecule in VEHICLe both singly and doubly, exploring every available exit-vector on each ring. The algorithm for the functionalising process is illustrated in pseudocode in Figure 3.2.

```
Data: List of carbon substituents, List of nitrogen substituents, List of
      VEHICLe molecules
Result: Library of monofunctionalised aromatic heterocycles

for molecule in VEHICLe molecules do
    identify exit-vectors;
    foreach exit-vector do
        Identify type of exit-vector;
        if exit-vector is C-H bond then
            for substituent in carbon substituents do
                Turn exit-vector into C-(substituent) bond;
                return C-functionalised molecule;
        else if exit-vector is N-H bond then
            for substituent in nitrogen substituents do
                turn exit-vector into N-(substituent) bond;
                return N-functionalised molecule;
```

Figure 3.2. The algorithm for functionalising the VEHICLe database.

The overall outline of the process illustrated in Figure 3.2 is as follows: each molecule in the VEHICLe database is considered in turn. For each molecule, the exit-vectors (functionalisable C-H or N-H bonds) are identified and a list made of these (in the Python implementation these are stored as lists of tuples of RDKit atom IDs, with the non-H atom ID listed first,

and the H atom ID listed second: `list([non-H atom ID, H atom ID])`). Each of these exit-vectors is then considered in turn, and the non-H atom in each exit-vector characterised as being either a carbon, or a 'pyrrole-like' nitrogen. If this atom is a carbon then each of the carbon substituents is 'bonded' to it to generate as many molecules as there are carbon substituents for the heterocycle being considered. If this atom is a 'pyrrole-like' nitrogen, then the process is the same as for carbon, only the nitrogen substituents are 'bonded' instead, thus generating as many new molecules as there are nitrogen substituents. In total, for each heterocycle the total number of monofunctionalised heterocycles generated by this process is $N_{\text{carbon exit-vectors}} \cdot n_{\text{carbon substituents}} + N_{\text{nitrogen exit-vectors}} \cdot n_{\text{nitrogen substituents}} + 1$, with $+1$ corresponding to the original unfunctionalised molecule, which is also returned. All of these created molecules are returned as SMILES strings, which are then canonicalised, assigned a unique RegID, and added to the database. Due to symmetries in many of the VEHICLe heterocycles, this process generates a number of identical molecules. As constitutionally-identical molecules share a canonicalised SMILES string, it was deemed more efficient to remove these after generation by matching identical SMILES strings than it was to check for symmetry in the heterocycles before functionalising. This was achieved trivially by casting the list of canonicalised SMILES strings as a set, thereby removing all redundancies, before assigning RegIDs.

The algorithm described above handles each molecule independently from the rest; the loops for identifying each molecule's exit-vectors and iteratively functionalising these are isolated within each molecule's scope. Furthermore the results are all SMILES strings, which are lightweight and easily and independently appended to an output list shared between processes. This makes these tasks embarrassingly parallel, thus lending itself to parallelised

processing. The functionalising of the 24 867 VEHICLe molecules was distributed across 8 CPU processors, and the entire database was monofunctionalised in under one hour.

To bifunctionalise the database, the process outlined in Figure 3.2 was repeated, using the monofunctionalised heterocycles as the input dataset. As the number of input molecules was larger by a factor of ten, the parallelisation provided significant efficiencies, and the dataset was bifunctionalised in under three hours. The total numbers of molecules in each dataset after functionalising is shown in Table 3.1.

Table 3.1. The number of heterocycles in each of the libraries after monofunctionalising the VEHICLe database, and bifunctionalising the monofunctionalised dataset.

Library	Number of Heterocycles
VEHICLe	24 867
Monofunctionalised	336 816
Bifunctionalised	486 220
Total	847 903

3.2.1 Substituent Selection

The criteria considered when selecting the substituents to use when functionalising the VEHICLe heterocycles are outlined below.

1. The substituents must be medicinally relevant, so groups that are known to appear regularly in approved drugs and chemical probes were considered. Groups that were known to be highly reactive or unstable were not considered.

2. The substituents must be small, with a limited number of conformational degrees of freedom. This is to avoid introducing the possibility of functionalised heterocycles having a large number of low-energy conformations, which would need to be considered

when aligning and scoring them. To avoid having to generate and minimise conformers (a computationally expensive task that would significantly increase the time taken to search a library), only small substituents were considered. Additionally, increasing the size of substituents (for example ethyl or isopropyl vs methyl) often yields diminishing returns in terms of electronic effects, while disproportionately increasing conformational flexibility and complexity.

3. No consideration was given to synthetic strategies for installing these substituents. As the purpose was to generate a complete enumeration of the region of chemical space defined by the starting library and these substituents, the concept of whether methodologies existed to synthesise each molecule was not considered at the point of creation of the expanded dataset. As the substituents chosen would be those that appear frequently in the medicinal chemistry literature, by definition synthetic strategies must exist to generate aromatic heterocycles with the selected functionality.

4. Acknowledgement was given that substituents bonded to aromatic carbons are likely to be different to those bonded to 'pyrrole-like' nitrogens, and as such different lists of substituents were selected for each.

Astex Pharmaceuticals published a recommendation engine in 2017 built using a graph network which included an analysis of the most popular ring substituents found in the ChEMBL database at the time of publication.[290] These substituents, and their number of observations in the ChEMBL dataset the authors analysed, are displayed in Table 3.2.

Using these as a starting point, and removing those that failed to meet the criteria outlined above (for example nitro was removed as it is a structural alert for toxicity, and the ethyl group was excluded as outlined in point 2 above), the final set of substituents for both carbon

Table 3.2. The top 10 ring substituents and their frequencies, as extracted from ChEMBL by Hall *et al.* 2017.[290]

Substituent	Frequency
	33 811
	24841
	18325
	12940
	11 124
	8591
	8583
	5984
	4200
	3229

and nitrogen were selected, and are outlined in Figure 3.3.[291] A trifluoromethyl group was included in the carbon substituents due to its popularity within medicinal chemistry, both as a means of blocking Phase 1 metabolism at reactive methyl sites and also as a means of increasing binding affinity.[292,293] *N*-methyl and *N*-trifluoroethyl groups were selected for functionalising at nitrogen.

Figure 3.3 panel C illustrates the functionalising process for indole. Each of the exit-vectors is indicated with a coloured circle. Each exit-vector is categorised as either being on carbon (green) or nitrogen (pink). All of the substituents shown in Figure 3.3 panel A are in turn bonded to each of the exit-vectors in green giving a total of 42 separate molecules, and each of those shown in Figure 3.3 panel B to the exit-vector in pink, giving a further two separate molecules. An additional 44 molecules for the searchable database are therefore generated for indole. This process is then repeated for each monofunctionalised molecule to generate bifunctionalised molecules.

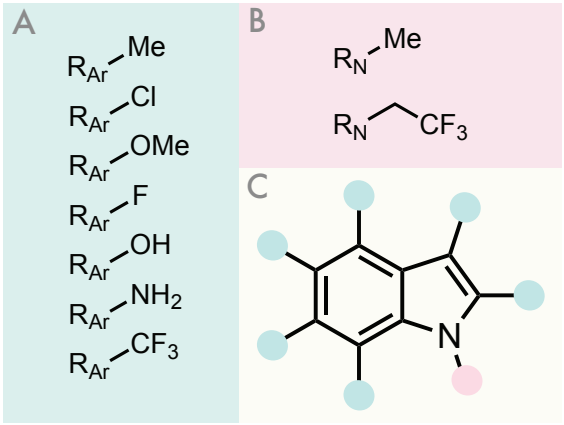


Figure 3.3. The substituents used to functionalise the VEHICLe database. **A** | The substituents bonded to carbon. **B** | The substituents bonded to 'pyrrole-like' nitrogen. **B** | An example of the functionalising process for indole.

3.2.2 Filtering

Although an objective of this thesis is to expand the region of synthesisable bioisosteric space, which necessitates the inclusion within the searchable library of molecules that have never previously been synthesised, included within the library generated using the methods described above are many molecules that could likely never stably exist. An illustration of these is given in Figure 3.4. Also included in the full database are molecules that, although medicinally reasonable, have no further aromatic exit-vectors, and thus cannot be further functionalised (**3.4** in Figure 3.4 illustrates this). Including these within the searchable database is not inherently problematic, as end users can exercise their judgement in determining which candidates to prioritise. However, their inclusion risks reducing the ranking of more sensible proposed bioisosteres. This increases the level of 'noise' in the returned results and may also undermine the perceived relevance of the results.

Simple filters were therefore implemented that would remove the molecules in the database that would be too unstable to exist or be medicinally useful (these are termed potentially explosive or bonkers, and are herein referred to as PEBs). The filters were chosen such

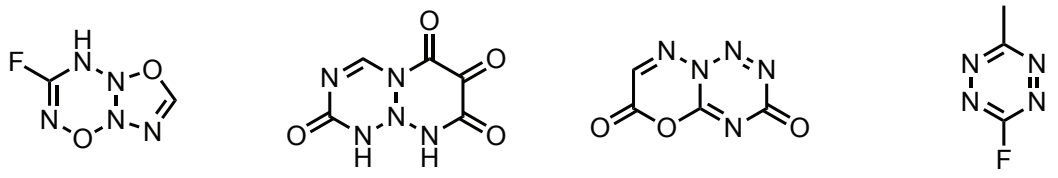


Figure 3.4. A illustration of physically unrealistic or medicinally impractical molecules present within the new dataset. **3.4** is included as, although chemically reasonable, it has no further aromatic exit-vectors, and as such presents no opportunity for incorporation into leads.

that they removed the minimum number of molecules, whilst leaving as many unsynthesised molecules in the searchable database as possible, and are outlined below.

1. Molecules with no further C-H or N-H exit-vectors, and are therefore not useful as bioisosteric candidates (for example **3.4**)

2. Molecules where the number of cyclic non-carbon heavy atoms is greater than the number of cyclic carbon atoms (for example **3.1**).

3. Molecules with four or more aromatic nitrogens bonded together (for example **3.2** or **3.3**). Tetrazole is manually excluded from this filter. This was implemented by removing molecules that match the SMARTS expression `n~n~n~n`.

4. Three or more carbonyls, using the SMARTS expression
   `'[R](=O).[R](=O).[R](=O)'` (**3.2** is an example of a molecule removed by this rule).

5. Any molecule with a cyclic anhydride, using the SMARTS expression
   `'[R](=O)o[R](=O)'`

6. Molecules containing a thioester, using the SMARTS expression `'[R](=O)s[R]'` .

These filters were applied to the combined database of unfunctionalised, monofunctionalised, and bifunctionalised heterocycles. The numbers removed, and the total numbers remaining after filtering, are listed in Table 3.3.

Table 3.3. Numbers of heterocycles removed due to lack of exit-vectors, matching with one or more of the filters, and the total remaining in each database, and the whole searchable library, after filtering.

Library	Total	No exit-vectors	PEBs	Remaining
Unfunctionalised	24 867	1476	13 262	10 129
Monofunctionalised	336 816	24 746	138 234	173 836
Bifunctionalised	486 220	40 091	83 823	362 306
Total	847 903	66 313	235 319	**546 271**

The three libraries were combined after filtering to give a final searchable database of 546 271 heterocyles, each identifiable with a unique RegID and a canonicalised SMILES string. Taken together these molecules form the MoBiVic database, which is used as the searchable database throughout the rest of this thesis (unless indicated otherwise).

3.3 Results and Discussion

It is interesting to note from Table 3.3 that the number of bifunctionalised molecules is not significantly higher than that of the monofunctionalised set. This is partly due to the increased likelihood of symmetry when adding a second substituent to a molecule already bearing one; functionalising the same scaffold with the same substituents in different orders or orientations can often lead to duplicate structures. Additionally, many monofunctionalised molecules possess only one or two available exit vectors, limiting the number of ways in which they can be further functionalised.

After filtering, the total MoBiVic library contains 546 271 molecules, which is more than 22 times larger than the original VEHICLe database. It would have been desirable to determine the proportions of MoBiVic that exist already within the scientific and patent literature, however at the time of writing the University of Oxford lacked API access to the Beilstein database (now commercialised by Reaxys™). The proportion of MoBiVic that has been synthetically accessed is likely smaller than that of VEHICLe. This is because functionalising a ring system that has not yet been synthesised results in molecules that are also, by definition, unsynthesised. However, the reverse is not necessarily true as functionalising a ring system that has been synthesised does not guarantee that the resulting molecule has been made.

To observe whether the distributions of any key properties differ when VEHICLe is mono- or bifunctionalised, the molecular weights, cLogP (as calculated in RDKit), number of hydrogen bond donors, acceptors, and heterocycles were calculated for each molecule in each of the three datasets. These data are plotted in Figure 3.5. As expected, the distribution of molecular weights is skewed towards higher weight as the library is functionalised. The calculated cLogP also tends towards higher lipophilicities as the datasets are further functionalised. This is expected when considering the substituents in Figure 3.3, as 6 of the 9 substituents generally tend to raise the lipophilicity of an aromatic molecule. The pattern in hydrogen bond acceptors initially appears surprising, with the unfunctionalised molecules tending to have a higher number of H-bond acceptors than the functionalised libraries. However, considering that many of the heterocycles in VEHICLe contain a large number of heteroatoms that can act as H-bond acceptors, and that these molecules (many of which may lack vectors for further functionalising) are then excluded from the other datasets, this distribution is less surprising.

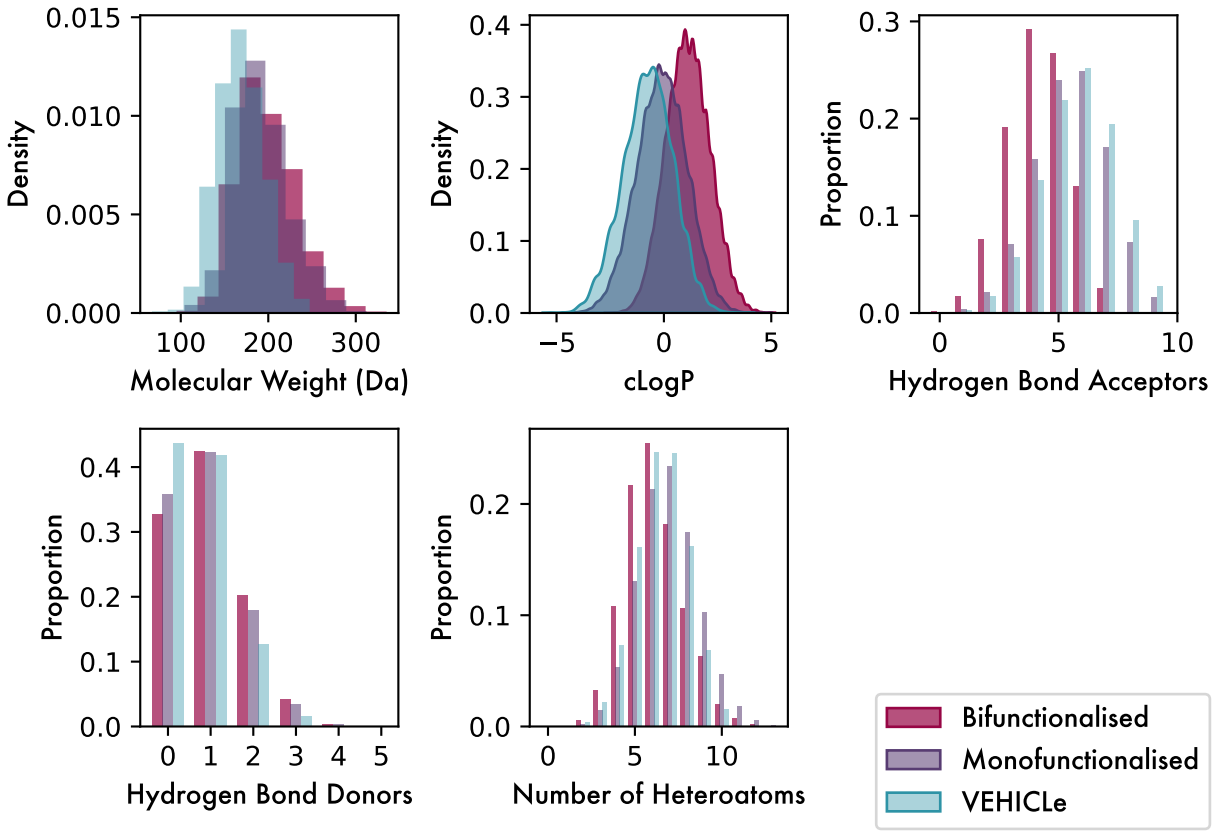


Figure 3.5. A comparison of the medicinally-relevant properties of the VEHICLe, monofunctionalised, and bifunctionalised datasets.

To observe the effect of removing the PEB-flagged molecules from the combined database of VEHICLe, mono-, and bifunctionalised molecules, the same parameters were calculated and plotted in Figure 3.6 for the combined datasets with and without PEB molecules. It is interesting to note that these distributions are near identical, however the distribution including PEB molecules has a slight hydrophobic tail. This is expected as the highly hydrophilic substituents (-OH and -NH_2) both add to the overall heteroatom count of a molecule, and as such are more likely to trigger the heteroatom-count PEB flag.

The inclusion of hydroxy- and amino-substituted heterocycles increases the number of structures in the database that are capable of tautomerism. Whereas VEHICLe includes only lactam tautomers, the addition of hydroxyl groups allows for the inclusion of corresponding

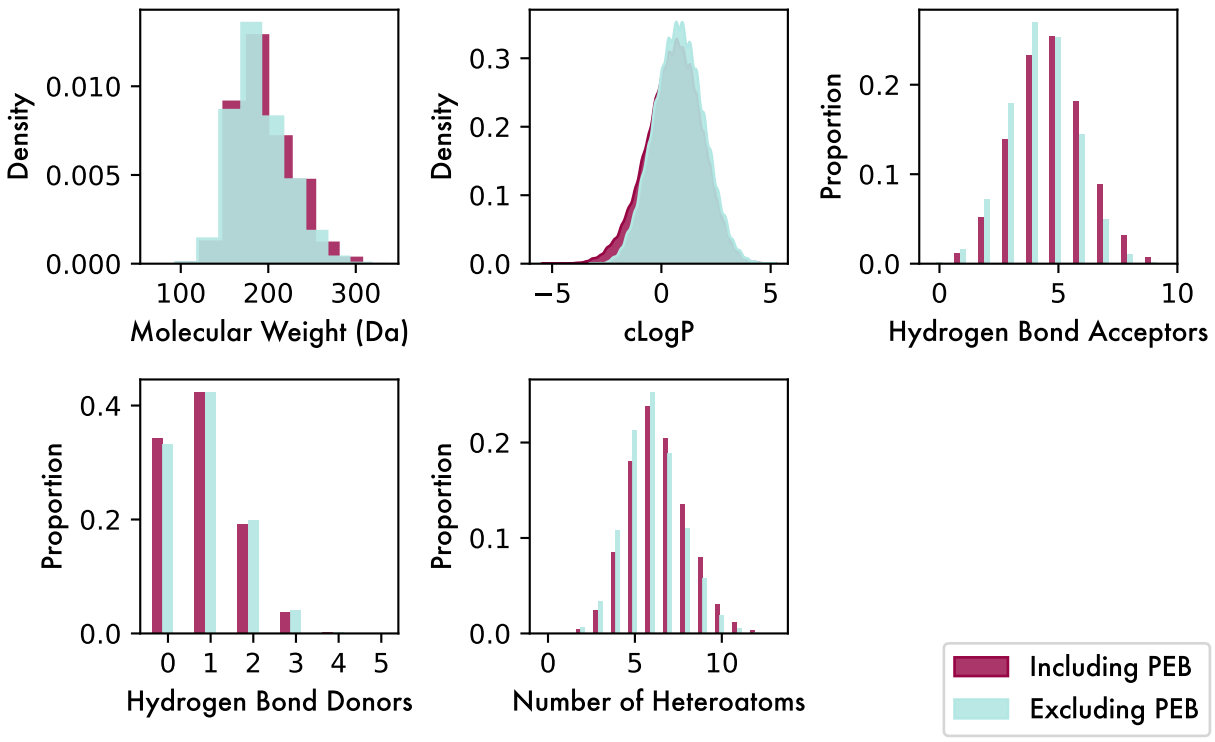


Figure 3.6. A comparison of the medicinally-relevant properties of the combined datasets with and without PEB molecules.

lactim tautomers, thereby expanding tautomeric coverage in MoBiVic. Amino-substituted molecules can exist as both amine and imine tautomers, though MoBiVic includes only the amine forms due to its construction constraints. As annular tautomers are enumerated for all relevant scaffolds, MoBiVic provides a sufficiently broad representation of tautomeric space for the identification of novel aromatic bioisosteres.

3.4 Conclusions

This chapter has described the expansion of the VEHICLe database by the sequential mono- and bifunctionalising of the VEHICLe heterocycles with a collection of medicinally-relevant substituents chosen based on a literature analysis of ChEMBL. This dataset was then further

filtered to remove likely unstable or medicinally unfeasible molecules. Each dataset is freely available on GitHub. This expanded dataset (MoBiVic) significantly expands the region of chemical space that is available for bioisostere searching, and thus increases the possibilities of discovering new aromatic heterocyclic bioisosteres of query molecules.

4 Development of the Current Implementation

4.1 Introduction

The implementation described in Chapter 2 (herein referred to as the first generation implementation) suffered from several shortcomings:

1. The searchable heterocyclic library contained only unsubstituted heterocycles, and the presence of heterocycles that are unlikely to be stable can skew the results.

2. There is no mechanism by which the user could specify an important attachment point, or several important attachment points, on the query molecule.

3. The underlying code used for the alignment is proprietary, and therefore has to be treated as a 'black box'.

The first of these shortcomings was resolved through the creation of the MoBiVic library in the previous chapter. Of those remaining the second is the most significant, as the substitution pattern of heterocycles in drugs is crucial in determining the strength of interaction with

the target protein.[284] However, as discussed in Section 2.2.3, the 'black-box' nature of the ShaEP code used for aligning and scoring means it is not possible to address this exit-vector issue without refactoring the alignment and scoring code completely.

This chapter describes the conceptual development and implementation in Python of a novel, vector-based alignment algorithm that aims to address the flaws with the implementation described in Chapter 2 and outlined above. The algorithm is extended to one user-specified vector and two user-specified vector searches, and moves away from the field-graph based electrostatic scoring method implemented in ShaEP. Efforts towards its benchmarking and preliminary results are described.

4.2 Methodology

This section describes the development and implementation of the computational approach to aligning and scoring molecules by exit-vector. It outlines the design of a unique exit-vector-based method of ligand alignment, and its implementation in Python. The scoring of a particular ligand pose relative to a query ligand, based on the similarity of the steric volumes of the ligands (the shape) and the similarity between each of the ESPs (electrostatic similarity) is discussed, and the implementation of these scoring functions in Python explained. As with the first-generation implementation, the basic workflow to return potential bioisosteric molecules from within a library of probes is as follows:

i. Each probe molecule in the library is aligned effectively to the query molecule.

ii. Each alignment is then scored to determine similarity.

iii. The highest scoring alignment for each probe is returned, and the results ranked in order of most similar to least similar

The objective of this tool is to take an aromatic ligand with one or two exit-vectors of significance highlighted by the user, and return in a useable format a list of possible bioisosteres of this molecule for use in medicinal chemistry campaigns. The approach developed here uses these exit-vectors as anchors to which each probe ligand is aligned and then shape and ESP similarity are used to score the alignment. All the aligned probes are then returned to the user, ranked in order of highest combined score to lowest combined score. The alignments of the top-scoring probe molecules are also available for inspection by the user.

In order to search any virtual database of molecules a software-based representation of a compound must first be created which encapsulates all the properties of the molecule relevant to the searching procedure, thus the representation of molecules as digital objects within computer software is discussed, and the object-oriented approach used here explained. The specification of a query molecule by the user as a SMILES string, and the use of placeholder atoms in this string to designate the desired attachment points follows from this. Once a digital representation of the query molecule and the probe molecules have been created, each probe molecule must be aligned in turn to the query molecule in a manner such that only alignments where functionalisable bonds (that is bonds that are amenable to conjugation to further chemical functionality in a synthetically feasible manner) align with the designated exit-vector(s) of the query are considered as valid for scoring.

The precise method of alignment differs depending on whether a single exit-vector or two exit-vectors is specified in the query by the user. For a single user-specified exit-vector, alignment is achieved by first identifying all the functionalisable bonds in each probe molecule (in this

case, any aromatic C-H or N-H bond from a 'pyrrole-like' nitrogen in a ring), and then using vector geometry to align these bonds along the designated query exit-vector sequentially. The alignment method also ensures that root mean square deviation (RMSD) between the planes of the rings is minimised. Every functionalisable bond in the probe is aligned and scored to the user-specified exit-vector in the query in a manner that takes into account any lack of rotational symmetry in either the query or the probe about the exit-vector of interest. Each of these alignments is scored, and the alignment with the highest total score (that is combined ESP and shape score) is returned, with the exit-vector in the probe that aligns with that of the query for this highest scoring alignment indicated. This is repeated for every probe in the library. This generates a list of scored probe molecules, with the exit-vector corresponding to the highest scoring alignment specified. These are sorted by total score, and returned to the user. Furthermore, the coordinates of a user-specified number of the top-scoring ligands in the highest scoring alignment are returned to the user for visualisation.

For two user-specified exit-vectors the searching method is somewhat different. An 8-bit binary hash is created for the query molecule, which is a compact identifier encoding the relative spatial arrangement and angular relationship of the two exit-vectors concisely. In order to avoid unnecessarily searching through probe molecules that do not have exit-vectors in the correct orientation, only those probes containing at least one pair of exit-vectors that share the same hash as the query molecule are aligned to. The exit-vector pairs that have the same hash as the query molecule are then aligned to the exit-vectors in the query in a method that minimises the RMSD similar to that for the one-vector case. Asymmetry is again accounted for, and the highest scoring alignment for each probe is returned, with the exit-vectors corresponding to this alignment designated. These are again sorted by total

score and returned to the user, alongside the aligned coordinates of the highest scoring probe molecules.

In the sections that follow, the technicalities of the design of this algorithm, its implementation in Python, and the approaches taken to validate and test this methodology are discussed in detail.[a] Together these approaches form a comprehensive computational methodology for searching through chemical space for predicted bioisosteric pairings of aromatic heterocycles, and facilitate the prediction of novel bioisosteres from both within and outside the synthesised region of aromatic heterocyclic chemical space.

4.2.1 Molecular Searching and Alignment

The algorithms developed for searching the library of probes and aligning each to the query depends on the number of vectors specified by the user, with the one-vector case handled separately to the two-vector case. It is true that heterocycles in medicinal chemistry can have more than two points of substitution; analysis of the ring systems present in molecules found either in FDA-approved drugs or clinical trials by Shearer *et al.* in 2022 found that rings in monocyclic compounds had on average 2.5 substitutions per ring (albeit with a large standard deviation), and bicyclic systems had on average one substitution per ring (thus an average bicyclic molecule would carry two substituents).[294] However, when coupled with Hall *et al.*'s analysis of ChEMBL ring substituents, it is likely that at least one of these ring substituents is one of those described in Section 3.2.1, and is thus already present in the MoBiVic library.[290] This means that (for the purposes of this implementation) these substituents can be considered as being a part of the molecule, rather than a variable position, and thus do not need to be included as a point of attachment in the query specification. It

[a]Software versions, packages, and computational resources are detailed in Section 8.1.

was felt that this caveat would handle the majority of cases for which a query molecule carried more than two substitutions, and thus only handling one-vector and two-vector searches in this tool was deemed an appropriate trade-off. In applications where a third attachment point is essential, it is possible for users to inspect the returned results and manually select those which have a functionalisable bond in an appropriate orientation for the third exit-vector.

This section outlines the process for representing molecules in digital space, aligning a user-specified query molecule to each of the probe molecules in the MoBiVic library, scoring each of these alignments according to shape and ESP similarity, and returning a list of proposed bioisosteres, ordered by similarity.

4.2.1.1 Molecule Representation

Python has extensive support for object-oriented programming. Molecules, which chemists naturally conceptualise as collections of atoms and bonds with associated properties (e.g., molecular weight, charges, and coordinates), align intuitively with an object-based structure. Representing molecules as Python objects encapsulates these properties as attributes and methods, providing a user-friendly interface that mirrors a chemist's intuition. This abstraction allows users to focus on molecular behaviour in scientific contexts, without being burdened by implementation or details of data storage.

In this implementation, molecules are represented as instances of the `hcie.Molecule` class. This class manages high-level functionality for calculating and storing key parameters required for alignment (*vide infra*). It also leverages the RDKit `Chem.Mol` class for efficient computation and access to attributes such as atomic element labels and indices. Follow-

ing the DRY principle,[b] reusing RDKit's well-optimised functionality avoids redundancy, simplifies the codebase, and benefits from the performance advantages of RDKit's C++ implementation.[295,296]

To maintain clear separation between the custom functionality of `hcie.Molecule` and RDKit's features, the implementation avoids inheritance. Instead, an instance of the RDKit `Chem.Mol` class is created during initialisation and accessed through the `self.mol` attribute. This approach ensures modularity and avoids the complexities of tight object relationships between the two classes.

Another crucial advantage of using the RDKit `Chem.Mol` class for molecule creation is its handling of atom indices. In digital representations of molecules, each atom (whether unique by symmetry or not) is assigned a unique index which acts as an absolute reference for that particular atom, and is unchanged by any operation that is carried out on the molecule (for example rotation, translation, or reaction). This invariant identifier for each atom is exceedingly useful when aligning molecules (see Section 4.2.1.6) as it allows the plane of the aromatic ring to be unambiguously defined. Different software packages handle the assignment of atom indices differently, and there is no set standard for the assignment, thus molecules generated in different programs may have different indices.

RDKit's method of assigning atom indices is illustrated in Figure 4.1. The schematic representation of triazine, as a user might draw in chemical drawing software is illustrated, and its corresponding SMILES string is shown. RDKit assigns zero-indexed atom indices according to the order of the SMILES string input given by the user, as demonstrated by the red subscript numerals, which show the mapping between the order in the SMILES string

[b] "Don't Repeat Yourself"

and the RDKit representation of the molecule below (note that these have been added for illustrative purposes, and do not form part of the SMILES string). SMILES strings are constructed with implicit hydrogens, and by default RDKit does not add explicit hydrogens to its representation of molecules, thus they have to be specifically added by the user. These are then numbered starting from the lowest-numbered hydrogen bearing atom, as demonstrated in Figure 4.1. In this implementation, atom indices are used to identify, track, and refer to atoms in any instance where this might be necessary (for example in specifying exit-vector atoms, or identifying the atoms that define the planes of aromatic rings).

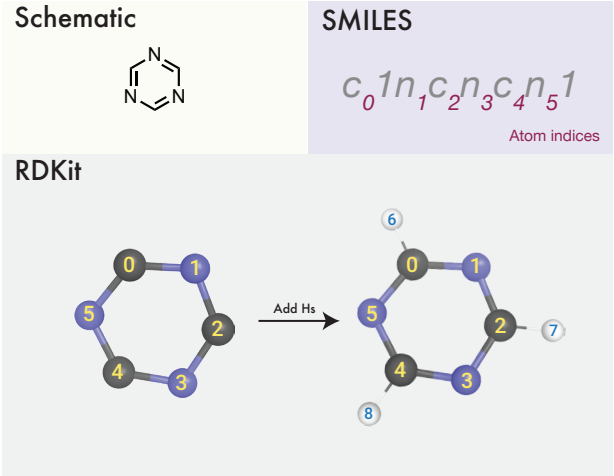


Figure 4.1. The relationship between molecular representations and atom indexing. The subscript numerals in the SMILES string demonstrate RDKit's assignment of atom indices, and are added for illustrative purposes only. These do not form part of the SMILES string.

4.2.1.2 Exit-Vector Specification

It was desirable to retain the SMILES input format for users in this implementation, as this is a universal, easily understood, and computer-readable molecular format that does not require the generation of complicated file structures and can easily be generated from graph-based molecular representations using commonplace chemical drawing software (for example ChemDraw™ or MarvinSketch™).

It was thus first necessary to indicate in the computer representation of the molecule which atoms are denoting the exit-vectors, and ensure that these are appropriately labelled with machine-recognisable dummy atoms[c]. Generic attachment points are not standardised within the SMILES language, and thus different software can generate non-canonical SMILES representations for the same molecule. For example ChemDraw™ designates attachment points within a SMILES string as `[R]`, but within RDKit generic attachment points specified in this way are not recognised. Thus in order to ensure that SMILES strings created in a wide variety of software packages are valid for input into this software, the initial step on instantiation of an `hcie.Molecule` is to identify any generic attachment points specified by non-supported SMILES characters in RDKit, and use the Python string replacement method to convert these to valid SMILES strings (a simple example is shown in Equation 4.1).

$$\underbrace{[R]c1ccccc1}_{\text{user input}} \longrightarrow \underbrace{[*:1]c1ccccc1}_{\text{RDKit recognised}} \tag{4.1}$$

When two vectors are specified, preservation of the labelling of these vectors is important for ensuring that the aligned probe molecules are returned with the corresponding exit-vectors correctly labelled. For example if the two exit-vectors in the query are labelled R_1 and R_2, then in each probe molecule the exit-vector closest to R_1 in the highest scoring alignment should be labelled with a 1 to indicate this. When an instance of an RDKit `Chem.mol` class is created from a SMILES string with two exit-vectors specified numerically (for example as `[R:1]c1cncc([R:2])c1`), the numerical labels on these attachment points are stored on each atom in the instance as `molAtomMapNumber` properties. This property is used in the

[c]In this context, dummy atoms refer to artificial placeholder atoms that bear no relation to actual elements and play no formal role in the chemistry of the molecule, except to demarcate points of attachment to further functionality.

`hcie.Molecule` class to set the order of the exit-vectors in the `self.user_vectors` attribute, which stores the atom indices of the exit-vectors specified by the user in the query as a list. As lists are ordered collections in Python, this order is preserved throughout the lifetime of the searching operations, and enables the correct labelling of the exit-vectors in the returned probe molecules.

4.2.1.3 Geometry Optimisation

The importance of geometry optimisation to the comparison of small-molecule ligands by shape similarity is described in detail in Section 2.1.3.1. However as the number of probe molecules in MoBiVic (> 500000) is more than 20 times larger than the original VEHICLe database, it is essential that the method chosen to optimise the geometry of the probe and query molecules is very fast. In order to minimise external dependencies, it is also preferable if the method of geometry optimisation does not depend on multiple external software packages.

RDKit has a built-in geometry optimisation algorithm that avoids expensive quantum mechanical calculations by using an experimental torsion-knowledge distance geometry (ETKDG) algorithm.[297] This algorithm generates realistic geometries for small molecules by initially creating a random estimate of the bond lengths in the molecule, using constraints derived from crystallographic data. These are then refined using experimentally-derived knowledge of bond lengths, bond angles, and dihedral angles until a set of criteria are met. An outline of the process is given below:

1. A bounds matrix is initialised.

   This is a matrix that determines the maximum and minimum distances between the

atoms in the molecule. Atoms that are bonded to each other are constrained by typical bond lengths between atoms of the same type (for example the C-C aromatic bond length constraints differ from the C-C aliphatic constraints). Non-bonded atoms are constrained by their van der Waals radii.

2. The bounds matrix is smoothed using a triangle-inequality algorithm.

   This step is necessary to ensure that the generated constraints are physically realistic. As the bounds are established by considering the pairs of atoms independently, it would be possible for a distance constraint to exist that violated the triangle inequality. For example, if the maximum distance A-B was set to 1.4 Å, and B-C 1.5 Å, without triangle-smoothing A-C could be set to 3.0 Å, which is physically impossible. Triangle-smoothing corrects these sorts of errors in the initial bounds matrix to ensure physically realistic and achievable geometries are returned.

3. A random distance matrix of intra-atomic distances is created.

   The distances here satisfy the bounds in the smoothed bounds matrix.

4. An initial, trial embedding of the molecule is created.

   3D coordinates for each of the atoms in positions that satisfy the bounds in the smoothed bounds matrix, and the distances in the random distance matrix, are created.

5. The coordinates of the molecule are iteratively refined.

   The coordinates are adjusted so that each bond length is as close to the centre of the bounds as possible, the torsional angles between atoms are within a set of limits derived from analysis of X-ray crystal structures in the Cambridge Structural Database, aromatic rings are planar, and the steric clash between non-bonded atoms is minimised. This process is repeated until either every generated geometry is within a certain

RMSD threshold, or a user-specified maximum number of iterations (in the case of this implementation this is set to 100 000) is reached.

This process is efficient and fast, generating reasonable geometries for a random selection of heterocycles from the MoBiVic library in an average of 3.7 ms per heterocycle.

Although the RDKit developers claim that the geometries generated from this algorithm are good enough for most computational purposes without further refinement, it is often standard practice to further optimise ETKDG-generated geometries with a molecular force field. Molecular force fields model the total energy of a molecule as the sum of various terms, each of which is approximated by a parametrised model. This is shown in Equation 4.2. In RDKit the MMFF94 (standing for Merck Molecular Force Field) is reliably implemented, and has been show to adequately refine small molecule geometries.[298] This approach again avoids the use of directly solving the Schrödinger equation, and as such is computationally cheap to implement.

$$E_{\text{total}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torsions}} + E_{\text{van der Waals}} + E_{\text{electrostatics}} \tag{4.2}$$

For the same random selection of heterocycles, adding a further refinement of the ETKDG geometries with an MMFF94 optimisation in RDKit added on average 420 µs to the time taken to generate each molecule, and as such it was decided that this step would be included in the molecule generation workflow.

In order to benchmark this approach, the geometries of molecules generated using this workflow were compared to the geometries of molecules optimised using higher-level density functional theory (DFT) optimisations. A random selection of 250 molecules were taken from the

VEHICLe database and geometries generated using the ETKDG method described above. These geometries were then further refined using the MMFF94 molecular force field described above, the semi-empirical xTB approach developed by Grimme *et al.*, and using DFT at a medium-level and high-level of theory.[261] The RMSDs of the geometries obtained through these approaches were then compared.

For the DFT calculations the PBE0 functional was chosen with a D3BJ dispersion correction, and the basis sets varied. For the medium-level calculation the def2-SVP (split valance polarised) basis set was used. This represents the valence orbitals of all atoms with two Gaussian functions with different exponents, and a single layer of polarisation functions are added to represent the distortion of atomic orbitals during bonding. This is a compact and computationally efficient basis set and has been shown in the literature to provide scientifically useful results for geometry optimisations, especially in systems with low levels of flexibility.[299] For the higher-level calculation the triple-zeta def2-TZVPP (valence triple-zeta with two sets of polarisation functions) was chosen. This approximates the atomic valence orbitals with three separate Gaussian functions, and adds two sets of polarisation functions to take into account orbital distortions during bonding. This is regarded in the literature as being the highest-level of theory required for accurate geometry calculations, but is a larger and more complex basis set than def2-SVP, and as such is more computationally demanding.[287,299,300] All calculations were carried out using the DFT implementations in the ORCA 5.0 software package using 4 CPU cores per calculation.[301]

The average time taken to optimise each molecule for each method is shown in Table 4.1. It is clear from these data that the DFT-based optimisation methods are significantly slower than the parametrised models implemented in RDKit, with the highest-level TZVPP basis set taking over one hour per molecule. The xTB approach is substantially faster than any of the

Table 4.1. The average time taken (per molecule) to optimise the molecular geometries across the various methods used in this benchmarking.

Method	Average time per molecule
ETKDG	3.7 ms
MMFF94	420 µs
xTB	246 ms
D3BJ-PBE0/def2-SVP	19 minutes
D3BJ-PBE0/def2-TZVPP	1 hour 6 minutes

ab-initio DFT approaches, but is still several orders of magnitude slower than the MMFF94 and ETKDG methods. To compare the geometries generated for each method, the RMSDs of the molecular geometries from each method were calculated from that generated by the initial ETKDG optimisation, and the histograms of these results are shown in Figure 4.2.

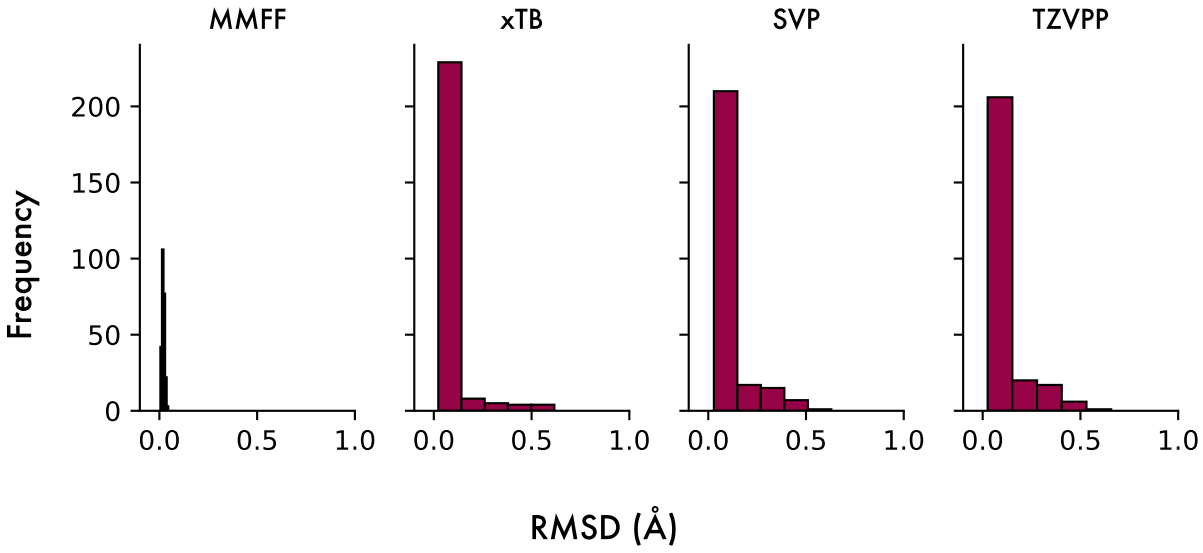


Figure 4.2. The RMSDs for the molecular geometries generated by each higher-level optimisation method, relative to the geometry generated by the ETKDG method in RDKit. For the DFT methods the D3BJ-PBE0 functional was used, and the label refers to the basis set used for the calculation.

These histograms show that the geometries after MMFF optimisation are most similar to those generated by ETKDG, with all of the RMSDs below 0.1 Å. The range of RMSDs is greater when the xTB and DFT methods are compared, with these geometries having

RMSDs between 0.0 – 0.6 Å. Pleasingly all of the RMSDs are well below 1.0 Å, and the large majority of the RMSDs for both the medium and high levels of theory are under 0.1 Å. In the development of the ETKDG algorithm, Riniker *et al.* use a 1.0 Å cut-off as a threshold for deciding whether conformers of the same molecule are different, and in molecular docking studies it is common to use 2.0 Å as an upper-bound for deciding if a crystal structure has been successfully reproduced.[297,302] Therefore having the majority of molecules < 0.1 Å across the three methods and no molecule > 0.6 Å is sufficient to conclude that the higher-level geometry optimisations offered by DFT are unnecessary in this context. Furthermore, the significant time penalty that these methods incur would render any software tool dependent on these less convenient to use. Although the xTB method is substantially quicker, it is still nearly 100 times slower than a simple ETKDG optimisation alone, and its results are not significantly different enough to warrant its use in this tool.

Additionally, the shape scoring method used in the final implementation (see Section 4.2.1.5) is relatively insensitive to small variations in atomic position, thus if the sum of all of the variations in atomic positions is < 1.0 Å, then these are unlikely to have any significant effect on the overall shape score between two molecules.

These results thus verify that an ETKDG geometry optimisation followed by an MMFF94 refinement are sufficient to generate reasonable geometries for further ligand comparison in a practical time-frame, and as such this method was implemented in the workflow.

4.2.1.4 Partial Charges

The concept of atomic partial charges and their significance in ESP scoring was discussed in Section 2.1.3.2. There the justification for the selection of xTB charges was given based

on an analysis of the correlations with bioactivity for a series of MPro inhibitors, where the xTB charges gave a correlation in good agreement with the more expensive DFT charges. However, as has been mentioned previously, the MoBiVic database contains over 20 times more heterocycles than VEHICLe. To ensure that end-users can freely extend the database or customise the method, no partial charge information was pre-calculated for the library, and as such the implemented method of calculating partial charges must be computationally efficient.

As the partial charges depend on the underlying representation of the electron distribution in the molecule, they are sensitive to the method used to calculate the geometry and electron distribution. Within RDKit is implemented a simple and efficient method for approximating partial charges using an iterative method of orbital electronegativity equalisation described by Gasteiger and Marsili in 1980.[303] This involves iteratively distributing charge based on the bond order and electronegativity difference between atoms in bonds, until the partial charges are stable across the molecule. It does not involve electronic structure calculations and is geometry independent and so is very quick, taking on average 11 µs for the molecules in the testing set. Although the results in Section 2.2.2.1 suggested that Gasteiger charges did not give the best correlations with bioactivity, in light of the need for a very rapid charge calculation, and as the scoring method is different in this implementation (*vide infra*), these Gasteiger charges were trialled as the partial charge method.

In order to benchmark the appropriateness of these simple charges, the charges calculated using the geometries and electronic structures obtained from the methods described above were compared to those calculated using the Gasteiger method. As partial charges are not a physical observable, there is no ground-truth with which to compare each method.

To compare these, each method's charges were plotted against those calculated with the Gasteiger method and the strength of the correlations observed.

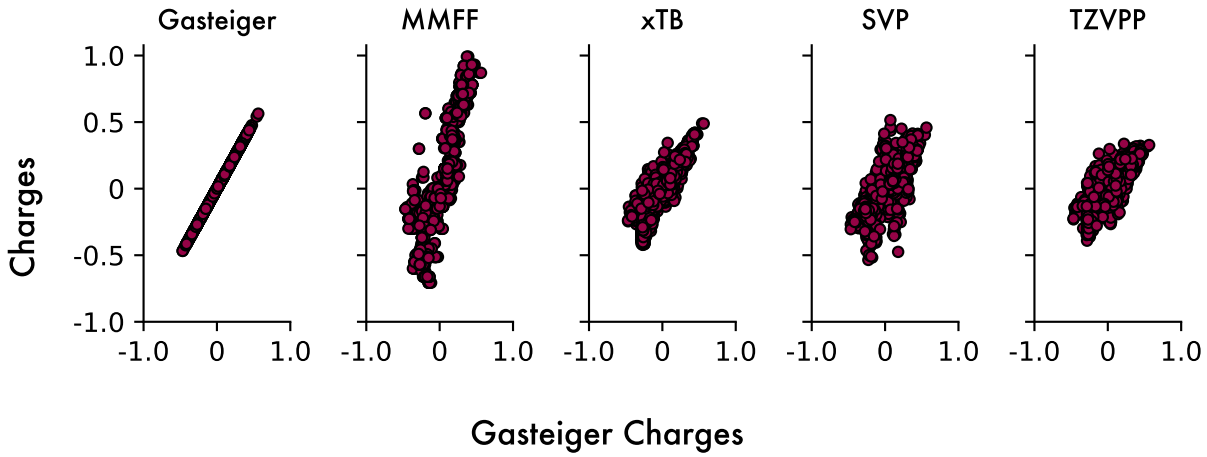


Figure 4.3. A comparison of the partial charges calculated with various electronic structure methods to those calculated by Gasteiger's method.

Figure 4.3 shows these correlations. It must be noted the precise methods for partitioning the charges differs between the methods, but this is inconsequential to the analysis here.[d] It is clear from these data that the general trend is one of strong positive correlation between all four methods analysed and the Gasteiger method. The range of charges is similar for all methods except that of the MMFF, which shows a much larger range of positive charges, including several above 0.9. This is physically unrealistic, as an atom with a partial charge this large indicates a bond with significant ionic character, none of which is expected in this dataset. The MMFF charges were therefore discounted from further consideration.

The correlations between the semi-empirical xTB method and the DFT-methods with the Gasteiger charges are all positive and generally strong, with the charges spanning a similar range across the four methods. It is interesting to note the difference in distribution between the charges calculated using the def2-SVP and def2-TZVPP basis sets, with the TZVPP

[d] xTB uses Hirshfeld surface analysis to partition the charges, whereas the DFT calculations in ORCA use Mulliken population analysis.

displaying a narrower distribution of charges to that of the SVP set. It is widely documented in literature that the Mulliken population analysis method used by ORCA in calculating these charges is sensitive to the basis set used in the underlying electronic structure calculation, and therefore the differences displayed here are not unprecedented.[304]

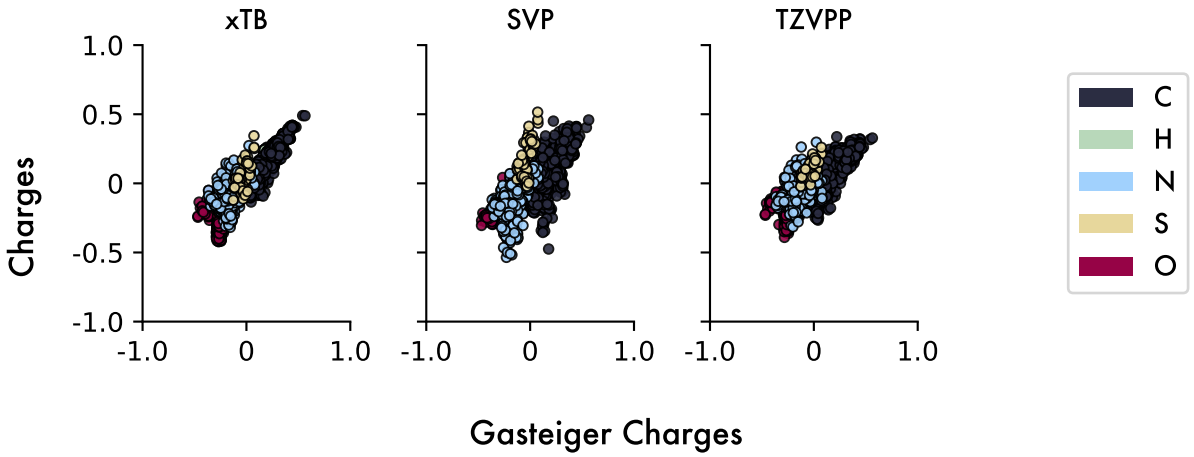


Figure 4.4. Charges by element

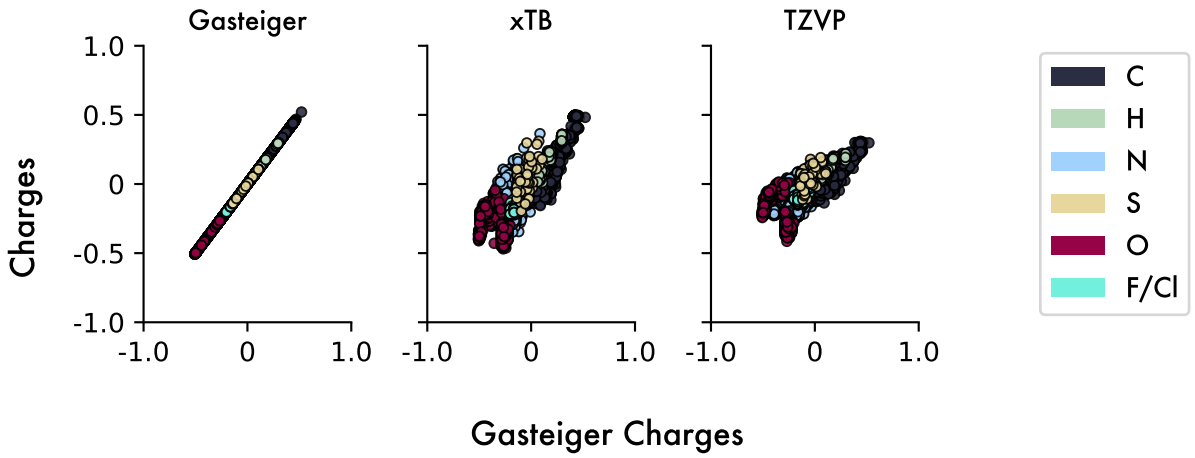
To investigate whether any deviations from the Gasteiger charges were due to differences in elemental parametrisation across any of the methods, the plots in Figure 4.3 were classed by element and are displayed in Figure 4.4. These show that the Gasteiger method almost exclusively assigns nitrogen, sulfur, and oxygen negative partial charges, whereas the other methods also assign these elements positive charges. As the Gasteiger method is solely dependent on bond order and relative atomic electronegativity, this is unsurprising. Other than this observation, Figure 4.4 offers no clear conclusions as to the origins of the differing charges.

Although it is clear that there are differences in the charges between the various methods, the similarity was deemed to be sufficiently good that the significant advantages in computational time gained by using the simple Gasteiger charges outweighed any gain in accuracy that would arise from a higher-level electronic structure calculation to determine these partial

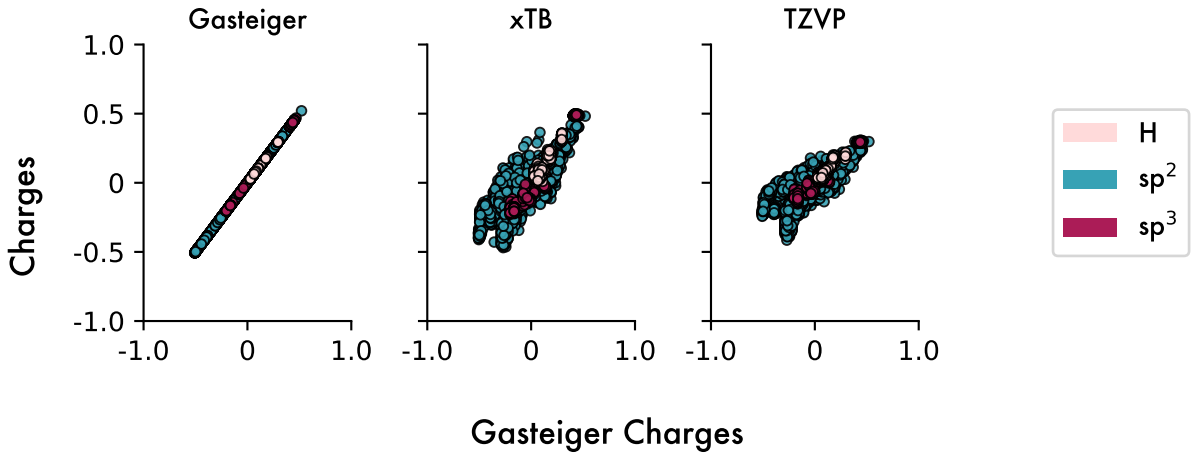
charges. Furthermore the avoidance of dependency on electronic structure packages improves the sustainability and maintainability of the codebase, which in turn aids in extending its usability to those with limited computational experience.

The benchmarking studies had thus far been on a sample of the original VEHICLe database. To ensure that the conclusions reached above were valid for MoBiVic, a random sample of 250 molecules taken from the expanded database were optimised using the ETKDG and MMFF method described above, and then these optimised geometries were used to calculate single-point energies (and thus partial charges) using the semi-empirical xTB method, and a DFT method. In order to balance accuracy with computational speed, the def2-TZVP (triple zeta valence polarised) basis set was used for the DFT calculation. This 'middle-tier' basis set captures a higher level of electronic structure detail than the simpler SVP set used above, but lacks the diffuse functions present in the TZVPP basis set, and is therefore faster to calculate.

Figure 4.5 shows the data from this analysis. It is interesting to note that the inclusion of substituents increases the range of charges calculated using the xTB method, but the DFT-based method retains a similar distribution to that of the original, unsubstituted sample. These charges were categorised by element in Figure 4.5a, but as with Figure 4.4 there appeared to be no obvious relationship between deviation from Gasteiger charge and element. As this dataset now included sp^3 hybridised atoms, any relation between hybridisation state and charge was investigated in Figure 4.5b. Although sp^3 hybridised atoms appear to be better correlated with the Gasteiger charges than the sp^2 hybridised aromatic atoms in both methods, there is little to suggest any fundamental relationship that would invalidate the use of Gasteiger charges.



(a) A comparison of the partial charges for a random subset of MoBiVic, coloured by element.



(b) Comparison of the partial charges for a random subset of MoBiVic, coloured by atomic hybridisation state.

Figure 4.5. Charge comparisons for a random subset of MoBiVic.

The average time taken per molecule to calculate these partial charges is displayed in Table 4.2. Although these represent a significant improvement on the times taken for geometry optimisation in Table 4.1, using xTB or a DFT-based method for partial charge calculation would still add a significant time overhead to searching the database of $> 500\,000$ molecules. As the benchmarking displayed in Figure 4.3 and Figure 4.5 does not indicate a significant difference between the Gasteiger charges and those calculated by more computationally ex-

pensive methods, the Gasteiger charges (as calculated in the RDKit implementation) were used for the remainder of this work.

Table 4.2. The average time taken per molecule to evaluate a single-point energy for the subset of 250 molecules randomly sampled from the functionalised database. The geometries used were those optimised by ETKDG followed by MMFF in RDKit.

Method	Average time per molecule
ETKDG	11 µs
xTB	341 ms
D3BJ-PBE0/def2-TZVP	6 minutes 56 s

4.2.1.5 Scoring

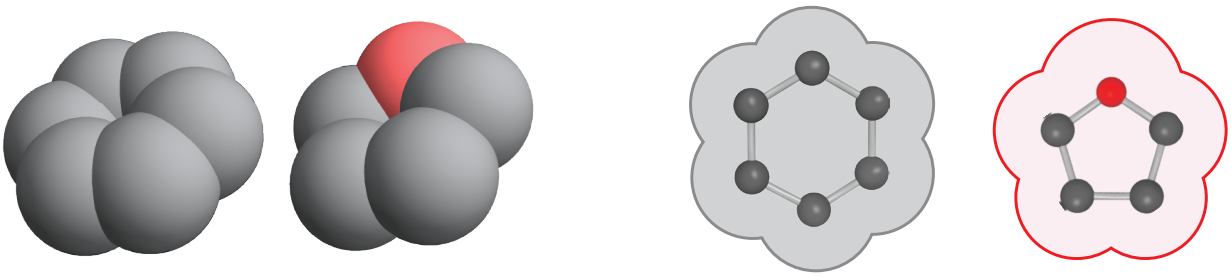
In order to quantify the similarity between molecules and thus rank proposed bioisosteres it is necessary to score the alignment between two ligands reliably and reproducibly. Each alignment is scored based on shape and ESP similarity, and the methods for determining these similarities are outlined below.

4.2.1.5.1 Shape Similarity

Quantifying the shape similarity between two molecules requires formalising the concept of a molecular shape. Crudely a molecule's shape is the region of space that it occupies, and is thus a function of the spatial coordinates (x, y, z). This means that each molecule's shape can be represented by a scalar field, defined in Equation 4.3.[305]

$$V(x,y,z) = \begin{cases} 1 & \text{if } (x,y,z) \text{ is within the molecular volume.} \\ 0 & \text{if } (x,y,z) \text{ is not within the molecular volume.} \end{cases} \tag{4.3}$$

There are various approaches to characterise the boundaries of a molecule's volume, including using Gaussian functions or wavefunction-based methods, however a simple approach involves using the van der Waals radius of each atom in the molecule to form a van der Waals surface, as illustrated in Figure 4.6.[306] Points within that surface are within the volume of the molecule, and those outside are not.[e]



(a) The van der Waals volumes for benzene (left) and furan (right).

(b) A horizontal section of the van der Waals volumes for benzene and furan. Hydrogens have been omitted for clarity.

Figure 4.6. The van der Waals molecular volumes for benzene and furan.

Two perfectly aligned, identical molecules would both have exactly the same scalar field. A molecule's volume is therefore just the sum of all of the positions in the scalar field that take a non-zero value:

$$\int V(x, y, z)dV \tag{4.4}$$

Quantifying the similarity in shape between two molecules A and B (for a given alignment) therefore can be reduced to comparing the volume of the region of the scalar field that is within both A and B to the volume that is in either A or B:

[e]The van der Waals radius arises from a hard-sphere model of atoms, and represents the distance of closest approach that another atom can make.

$$\text{Similarity} = \frac{\overbrace{\int V_A(x,y,z) \cdot V_B(x,y,z)\, dV}^{\text{Intersection of molecular volumes}}}{\underbrace{\int V_A(x,y,z)\, dV + \int V_B(x,y,z)\, dV - \int V_A(x,y,z) \cdot V_B(x,y,z)\, dV}_{\text{Union of molecular volumes}}}$$

$$= \frac{V_A \cap V_B}{V_A \cup V_B} \tag{4.5}$$

The intersection is subtracted from the sum of the two volumes in the denominator to avoid double counting. This metric is the Tanimoto similarity for shape similarity, and its components are illustrated for the simple case of a (poor) alignment between benzene and furan in Figure 4.7.[307] This takes values within the interval $[0, 1]$, with two molecules sharing no overlap scoring 0, and two identical molecules in perfect alignment scoring 1. This is typically calculated with reference only to heavy atoms, as these overwhelmingly dominate the molecular volume and thus hydrogen atoms are ignored for efficiency.

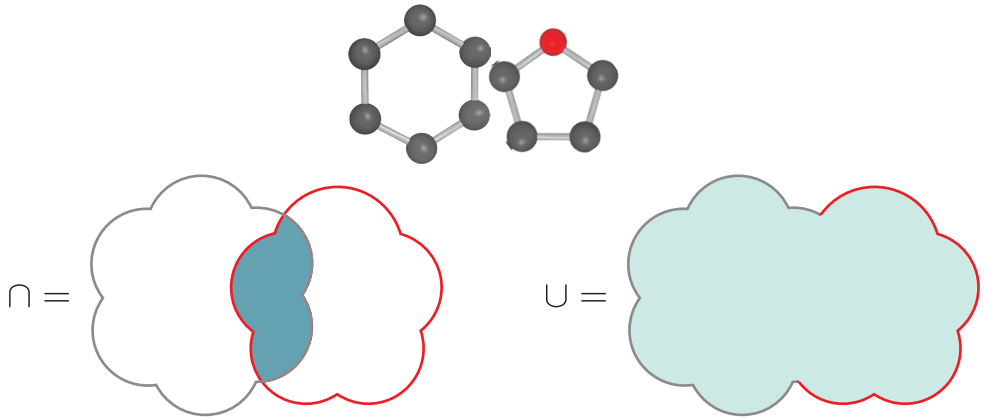


Figure 4.7. Calculating the Tanimoto shape similarity based on volume overlap. For clarity, the volume overlaps are illustrated with horizontal sections.

In this implementation, these are calculated for a given molecular alignment using RDKit's in-built function `AllChem.ShapeTanimotoDist`, with the similarity equal to 1 − the Tanimoto distance. The underlying source code is implemented in C++, thus this is a very quick and efficient way of calculating the Tanimoto shape similarity for a given alignment.

4.2.1.5.2 ESP Similarity

The method of scoring an alignment for ESP similarity used here is based on that first described by Good *et al.* in 1992, and implemented in Python by Bolcato *et al.* in 2022.[273,308] The electrostatic similarities are calculated based on the overlap of the electrostatic potentials of the two aligned molecules.

The value of the electrostatic potential at a point in space $\vec{r}$ for a molecule of n atoms, each with partial atomic charge q_i, is given by Coulomb's law

$$\phi_E(\vec{r}) = \sum_{i=1}^{n} \frac{q_i}{|\vec{r} - \vec{R_i}|} \tag{4.6}$$

where $\vec{R_i}$ is the position in space of the atom with partial charge q_i. The similarity between two molecules **A** and **B** can be calculated using a slight modification of the Tanimoto similarity index used for the shape similarity in Equation 4.5

$$\text{Similarity} = \frac{\int \phi_A(\vec{r}) \cdot \phi_B(\vec{r}) dV}{\int \phi_A(\vec{r})^2 dV + \int \phi_B(\vec{r})^2 dV - \int \phi_A(\vec{r}) \cdot \phi_B(\vec{r}) dV} \tag{4.7}$$

The squaring of the integrand in the self-overlap integrals of denominator of Equation 4.7 is necessary as the electrostatic scalar fields can take continuous values, rather than those that appear in the shape similarity calculation which can only be 1 or 0.

Directly integrating over Equation 4.6 can be problematic. As $\vec{r}$ approaches $\vec{R_i}$ the expression becomes singular, leading to numerical instabilities close to the atomic positions. Furthermore, directly integrating $\frac{1}{(\vec{r}-\vec{R_i})}$ over space is computationally costly due to the need to directly calculate pairwise interactions for each atom. To avoid these issues Good *et al.* approximated the $\frac{1}{r}$ dependency with a sum of three Gaussian functions whose coefficients had been selected to mimic the desired behaviour over the range of molecular interest (as demonstrated in Figure 4.8):

$$\frac{1}{r} \approx 0.3001 \cdot e^{-0.0499 \cdot r^2} + 0.9716 \cdot e^{-0.5026 \cdot r^2} + 0.1268 \cdot e^{-0.0026 \cdot r^2} \tag{4.8}$$

This eradicates the issue of singularity as $|\vec{r} - \vec{R_i}| \to 0$ as the Gaussians take a finite value at $\vec{r} = 0$. It is also significantly more computationally efficient as the product of Gaussian functions centred on different atoms is itself a Gaussian function, and a closed-form analytical expression for its integral exists and is computationally cheap to evaluate. A derivation of the closed-form expression for the overlap integral of these two-centre composite Gaussian functions is given below.

$$G_k^i = \gamma_k \cdot e^{-\alpha_k(\vec{r}-\vec{R_i})} \tag{4.9}$$

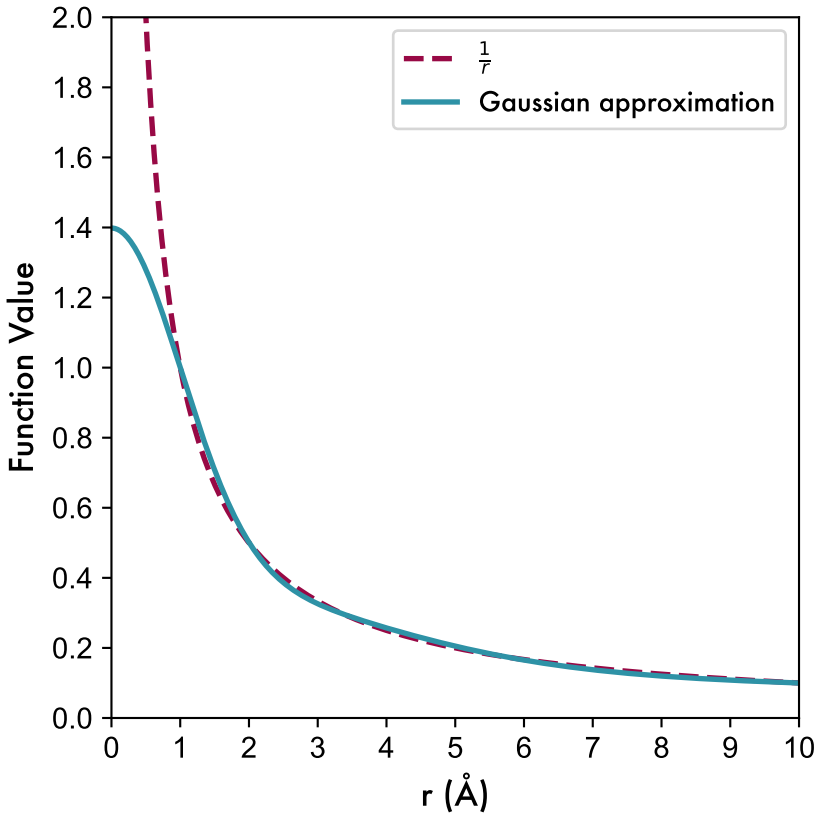


Figure 4.8. The Gaussian approximation to the 1/r curve, showing the singularity as $r \to 0$.

Substituting the three-Gaussian approximation 4.8 into 4.7 using the shorthand notation given in 4.9 gives

$$
\begin{aligned}
I_{AB} &= \sum_{i=1}^{n}\sum_{j=1}^{m} q_i q_j \int (G_1^i + G_2^i + G_3^i) \cdot (G_1^j + G_2^j + G_3^j) dV \\
I_{AA} &= \sum_{i=1}^{n}\sum_{i=1}^{n} q_i q_i \int (G_1^i + G_2^i + G_3^i)^2 dV \\
I_{BB} &= \sum_{j=1}^{m}\sum_{j=1}^{m} q_j q_j \int (G_1^j + G_2^j + G_3^j)^2 dV \\
\text{Similarity} &= \frac{I_{AB}}{I_{AA} + I_{BB} - I_{AB}}
\end{aligned}
\tag{4.10}
$$

Expanding the expression for I_{AB} gives

$$\begin{aligned} I_{AB} &= \sum_{i=1}^{n}\sum_{j=1}^{m} q_i q_j \int \sum_{k=1}^{3}\sum_{l=1}^{3} G_k^i G_l^j dV \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{3}\sum_{l=1}^{3} q_i q_j \gamma_k \gamma_l \int e^{-\alpha_k |\vec{r}-\vec{R}_i|^2} e^{-\alpha_l |\vec{r}-\vec{R}_j|^2} dV \end{aligned} \tag{4.11}$$

The Gaussian product in the integral can be rewritten as

$$e^{-\alpha_k |\vec{r}-\vec{R}_i|^2} e^{-\alpha_l |\vec{r}-\vec{R}_j|^2} = e^{-\frac{\alpha_k \alpha_l}{\alpha_k+\alpha_l}|\vec{R}_i-\vec{R}_j|^2} \cdot e^{-(\alpha_k+\alpha_l)|\vec{r}-\vec{R}_p|^2}$$

where

$$\vec{R}_p = \frac{\alpha_k \vec{R}_i + \alpha_l \vec{R}_j}{\alpha_l + \alpha_k}$$

Using the standard integral for a normalised Gaussian over all space:

$$\int e^{-(\alpha_k+\alpha_l)|\vec{r}-\vec{R}_p|^2} = \left(\frac{\pi}{\alpha_k+\alpha_l}\right)^{\frac{3}{2}}$$

enables 4.11 to be expressed as a closed-form sum over constants, atomic distances, and charges

$$I_{AB} = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{3}\sum_{l=1}^{3} q_i q_j \gamma_k \gamma_l \left(\frac{\pi}{\alpha_k+\alpha_l}\right)^{\frac{3}{2}} \cdot e^{-\frac{\alpha_k \alpha_l}{\alpha_k+\alpha_l}|\vec{R}_i-\vec{R}_j|^2} \tag{4.12}$$

In this implementation (based heavily on that described by Bolcato *et al.* in 2022), these integrals are calculated in matrix form, leveraging the efficiencies of the `numpy` and `scipy` libraries for matrix multiplication.[308] Each integral calculation in its matrix representation is

$$I = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i q_j \sum_{k=1}^{3} \mathbf{A}_k \cdot e^{-\mathbf{B}_k \cdot r_{ij}^2} \tag{4.13}$$

where

$$\mathbf{A}_{kl} = \gamma_k \gamma_l \left(\frac{\pi}{\alpha_k + \alpha_l} \right)^{\frac{3}{2}} = \begin{bmatrix} 15.906 & 3.953 & 17.615 \\ 3.953 & 5.216 & 1.910 \\ 17.615 & 1.910 & 238.758 \end{bmatrix}$$

$$\mathbf{B}_{kl} = -\frac{\alpha_k \cdot \alpha_l}{\alpha_k + \alpha_l} = \begin{bmatrix} 0.025 & 0.045 & 0.002 \\ 0.045 & 0.251 & 0.003 \\ 0.002 & 0.003 & 0.001 \end{bmatrix}$$

and r_{ij} is the Euclidean distance between atoms i and j.

In order to avoid erroneous ESP similarity scores caused by poor alignments of X-H hydrogens (where X $\in$ {O, N, C} and is not aromatic), these hydrogen atoms are excluded from the similarity calculations. A full justification for this, and an example of why it is necessary, is given in the Appendix.

Unlike the Tanimoto similarity for shape, 4.7 gives values in the range $[-\frac{1}{3}, 1]$.[f] To correct for this the final value is normalised to sit within the interval $[0, 1]$.

The final similarity score is the simple sum of the shape and ESP similarity scores, and sits in the interval $[0, 2]$. This score was not normalised to $[0, 1]$ to highlight that the total is a combination of shape and ESP scores. It is possible for the end-user to unequally weight the contributions of shape and ESP score to the total if desired.

[f] Conversations with Dr Esther Heid were invaluable in understanding this.

4.2.1.6 One-Vector Alignment

When a single exit-vector is specified by the user, the query molecule is aligned and scored against every probe molecule in the database using a vector-based alignment method. For each of the molecules in the searchable database the possible exit-vectors are enumerated and stored as a list of tuples of atom indices in the order `[non-H atom index, H atom index]`. The number of exit-vectors for each probe molecule is also stored as a property in the database.

The algorithm for searching the functional database for a single-vector is illustrated in pseudo-code in Figure 4.9 below. Every molecule in the searchable database is compared against the query molecule. For each probe molecule, all the exit-vectors are identified (this is trivial as the number of exit-vectors and their atom indices are stored as a property of each probe molecule in the searchable database) and twice the number of conformer objects as there are exit-vectors are created. Each exit-vector is then aligned sequentially to the user-specified exit-vector on the query molecule. The probes are aligned such that the RMSD between the exit-vectors and ring planes is minimised (*vide infra*). This alignment is then scored, and the coordinates of the alignment stored as a conformer object. The score is stored as an attribute of the instance of the probe molecule such that each score is linked to the conformer object storing the coordinates of the alignment. After the first alignment for each exit-vector is scored, the plane of the ring is then rotated by 180° about the axis of the exit-vector and scored again, with this score and the coordinates of the alignment stored in the subsequent conformer object. This rotation takes into account the potential lack of C2 symmetry about the exit-vector in question. Although there exist methods for ascertaining whether this symmetry exists, it was deemed to be faster simply to rotate and re-score about each exit-vector.

```
Data: Searchable library of aromatic heterocycles
Result: Heterocycles ranked in order of similarity

for molecule in searchable library do
    identify exit-vectors;
    foreach exit-vector do
        Align to query exit-vector;
        Align ring planes;
        Score alignment and store score;
        Rotate alignment by 180° about axis of exit-vector;
        Score alignment and store score;
    return Highest-scoring exit-vector alignment;
Rank probe molecules by total score;
Output to user;
```

Figure 4.9. The algorithm for aligning and scoring probe molecules in the searchable database against a query molecule with a single user-specified exit-vector.

Once this process has been completed for each exit-vector in the probe, the top scoring alignment and its respective total, shape, and ESP scores are stored in a list. These are collected for all of the probe molecules in the searchable database, and are then sorted by total score and returned to the user. The exit-vector of the highest-scoring alignment is indicated in the SMILES string returned to the user for each probe molecule, and the top 50 alignments are returned in an `sdf` file, displaying the molecular coordinates. This process is demonstrated visually for furan and oxazole in Figure 4.10 below.

The method of aligning each probe to the query is adapted from that first described by Kabsch in 1976, and involves constructing a matrix of the atomic coordinates of the query and the probe, and then using linear algebra techniques to find the rotation matrix of the probe onto the query that minimises the RMSD between their respective coordinates.[309,310] When the exit-vector for the probe molecule has been identified, the non-H atom (herein referred

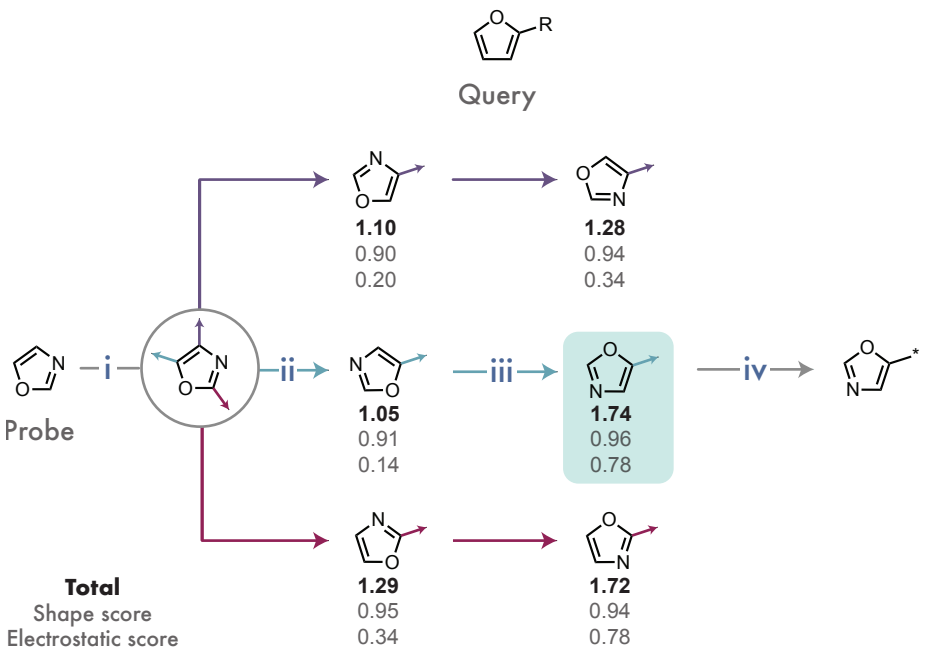
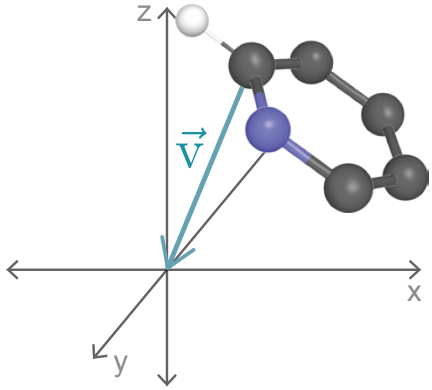
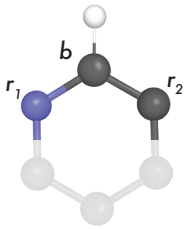


Figure 4.10. An illustration of the one-vector alignment and scoring process for an oxazole probe aligned to a furan query. **i** | All exit-vectors on the probe are identified. **ii** | Each exit-vector in the probe is aligned to the user-specified vector in the query and this alignment scored. **iii** | The alignment is rotated by 180° about the axis of the probe exit-vector to account for any asymmetry and re-scored. **iv** | The highest-scoring alignment is identified and returned to the user, with the exit-vector clearly indicated.

to as the *base* atom) is defined as the centre of rotation. The probe and query molecules are translated such that the base atom lies at the origin, as illustrated in 4.11a. As only two vectors are needed to define the plane of an aromatic ring, the matrix of coordinates required for calculating the rotation can be reduced to just three of the ring atoms. Centering the base atom (rather than the centroid of the molecule, as is usual when performing a Kabsch rotation) at the origin avoids the need to include the H-atom of the exit-vector (herein referred to as the *tail* atom) in these matrices. These matrices are constructed as shown in 4.11b, with the atoms of the probe forming the **P** matrix, and the atoms of the query forming the **Q** matrix.



(a) The vector ($\vec{v}$) defining the translation of the base atom to the origin.



$$\mathbf{P} = \begin{bmatrix} \longleftarrow & \vec{\mathbf{b}} & \longrightarrow \\ \longleftarrow & \vec{\mathbf{r}_1} & \longrightarrow \\ \longleftarrow & \vec{\mathbf{r}_2} & \longrightarrow \end{bmatrix}$$

(b) The construction of the **P** matrix for the 2-pyridine exit-vector. **b** defines the base atom of the exit-vector, and $\mathbf{r}_1$ and $\mathbf{r}_2$ are the ring atoms that define the plane of the ring. Their coordinates make up the rows of the **P** matrix.

Figure 4.11. Setting up the Kabsch alignment for 2-pyridine.

The optimal rotation matrix that rotates the points defined in **P** onto the points defined in **Q** such that the RMSD between them is minimised is then found by first determining the co-variance matrix

$$\mathbf{H} = \mathbf{P}^{\mathrm{T}}\mathbf{Q}$$

which captures the degree of alignment and spatial correlation between the points in each set. This is then deconstructed into its component matrices using a singular value decomposition

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$$

where **U** and $\mathbf{V}^{\mathrm{T}}$ are orthogonal matrices representing rotations, and $\mathbf{\Sigma}$ is a diagonal matrix representing a scaling. Finally, the matrix that represents the optimum rotation is found

$$\mathbf{R} = \mathbf{V}\begin{pmatrix} 1 & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 1 \end{pmatrix}\mathbf{U}^{\mathrm{T}} \quad \text{where } d = \det(\mathbf{V}\mathbf{U}^{\mathrm{T}}) = \begin{cases} +1 & \text{if rotation} \\ -1 & \text{if reflection} \end{cases}$$

The single value decomposition can, depending on the precise nature of **P** and **Q**, produce a matrix that corresponds to a reflection rather than a rotation. The inclusion of d in the identity matrix multiplication corrects for this occurrence.

This rotation matrix **R** is then applied to the full matrix of the probe's coordinates, to generate an alignment to the query $\mathbf{P}_{\text{aligned}}$, which is then scored as described above. $\mathbf{P}_{\text{aligned}}$ is then multiplied by a matrix corresponding to a 180° rotation about the axis defined by the probe exit-vector, and re-scored.[g]

This process is repeated for every exit-vector in each probe molecule, and the highest total score for each probe molecule is collected, ordered, and provided to the user as described above.

4.2.1.7 Two-Vector Alignment

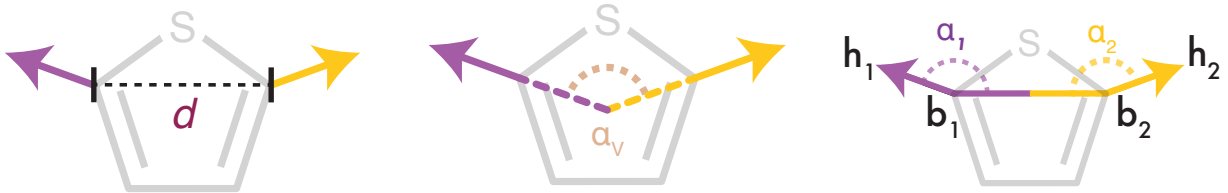
The process of aligning and scoring probes to a query where two exit-vectors are specified by the user is different. The geometry of the two exit-vectors is categorised, and only aligned and scored against probe molecules which have at least one pair of exit-vectors with a similar geometry. These geometries therefore need to be categorised in a manner that represents the arrangement of the exit-vectors whilst making the retrieval of molecules from the database with similar exit-vector arrangements fast and straightforward.

Inspiration was taken from the CAVEAT software published by Lauri and Bartlett in 1994, where a method of binning exit-vector pairs based on the angles and distances between them was described.[240] The overall principle is that a pair of exit-vectors can be characterised by

[g]This rotation matrix is calculated using the Euler-Rodrigues formula. A full description is outside the scope of this thesis.

a set of parameters, the values of which determine which 'bin' the arrangement is classified into. All pairs of exit-vectors in each molecule in the database are classified and assigned a bin. When a user specifies a pair of exit-vectors in a query, the bin corresponding to the specified arrangement is identified, and the query molecule is aligned and scored only against molecules in the same bin.

The original authors define four parameters to characterise an exit-vector pair, but two of these involve dihedral angles and are not applicable to aromatic systems. Consequently, only two parameters are used in this implementation to describe the vector pairs in the MoBiVic library, and these are illustrated in Figure 4.12. The distance between the two base atoms (labelled as $\mathbf{b}_1$ and $\mathbf{b}_2$ in Figure 4.12c) of the vector pair (Figure 4.12a) and the angle between them (Figure 4.12b) describe their geometry. The distance between the base atoms is calculated trivially from their coordinates, and the angle between the two exit-vectors is calculated using simple geometry ($\alpha_v = \angle \mathbf{h}_1\mathbf{b}_1\mathbf{b}_2 - (180 - \angle \mathbf{h}_2\mathbf{b}_2\mathbf{b}_1)$).



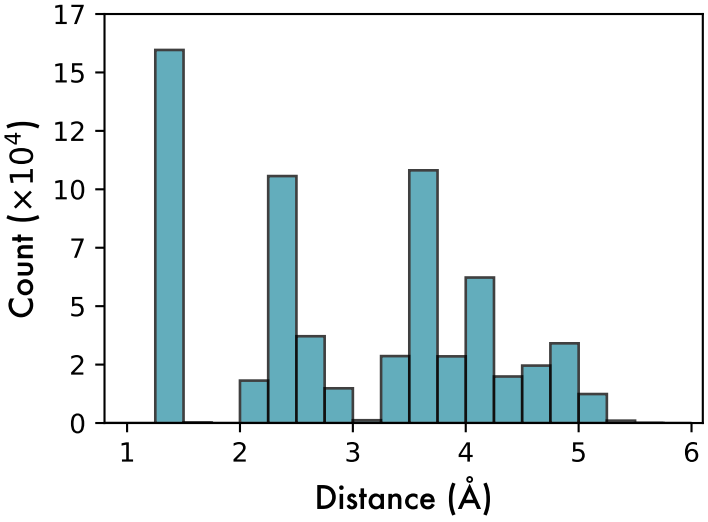
(a) The distance between the base atoms of an exit-vector pair.

(b) The angle (α_v) between the exit-vectors.

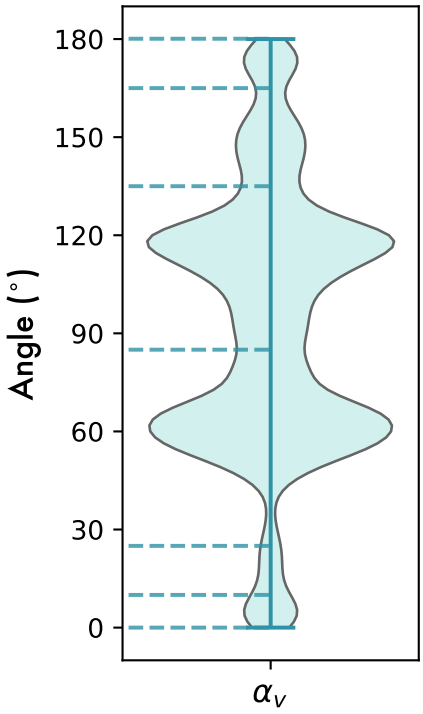
(c) The two angles used to calculate α_v for the same exit-vector pair.

Figure 4.12. The parameters used to characterise the geometry between a pair of exit-vectors (**a** and **b**), and the angles used to calculate the angle between the exit-vectors (**c**)

These parameters were calculated for each pair of exit-vectors in every molecule in the VEHICLe database and their distributions examined to determine the appropriate boundaries for each bin. These are displayed in Figure 4.13.



(a) A histogram of the distances between base atoms in the VEHICLe database.



(b) The angles between exit-vector pairs in the VEHICLe database. The horizontal lines indicate the bin boundaries.

Figure 4.13. The distributions of two-vector parameters for the expanded database of heterocycles.

As the arrangements of aromatic heterocycle exit-vectors are restricted by the geometry of the heterocycles themselves, it is not surprising that Figure 4.13 shows clear boundaries for discretisation. Based on the distribution in 4.13a, the distances were grouped into those where $d < 2.0$ Å, $d > 6.0$ Å, and bins of equal 0.25 Å width for $2.0 \leq d \geq 6.0$. The angle bins were as shown in 4.13b. Although for the angles around 90° there is no obvious bin boundary, the precise position of the angle boundary is not likely to end up limiting the returned results as the distances will likely act as a discriminator. Tables showing the precise boundaries and binary labels for each bin are shown in Appendix 9.3.

To encapsulate both the information about the distance and the angle for each exit-vector pair in a succinct and searchable manner, an 8-bit hash is created for each exit-vector pair geometry. To determine this hash each bin is ordered from smallest-valued to largest-valued, and the bins assigned a binary number based on this order (the bin containing the smallest numbers is assigned 0). The bins for distance and for angle are ordered and numbered separately. For the 18 distance bins, a 5-bit binary number is assigned to each bin, and for the 6 angle bins a 3-bit binary number is assigned. The distance bin binary label and the angle bin binary label are then combined to form an 8-bit binary hash for each exit-vector pairing, as shown in Figure 4.14.

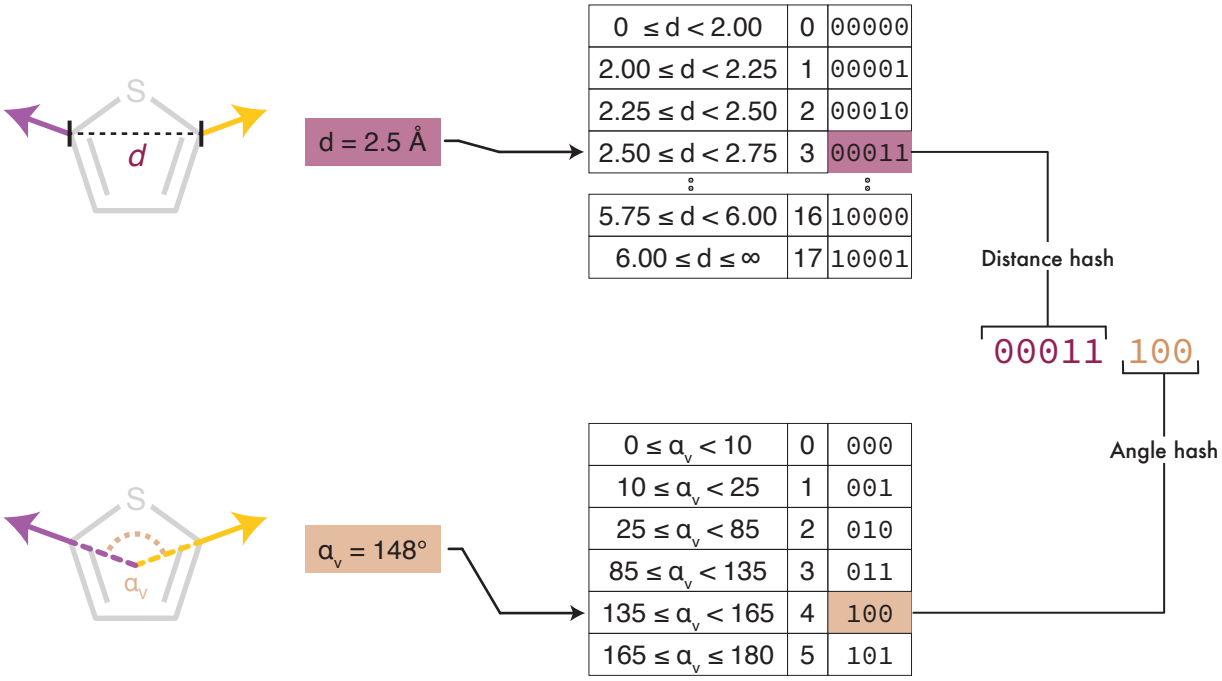


Figure 4.14. The process of constructing the 8-bit binary hash for each exit-vector geometry, as illustrated for thiophene.

A hash was then calculated in this manner for every vector pair in each molecule in the expanded database, as demonstrated for thiophene in Figure 4.15. Each exit-vector pair is identified, and the necessary geometric constants (d and α_v) calculated and a hash value

assigned as described above. This results in $\binom{n}{2}$ hashes per molecule, where n is the total number of exit-vectors. It is interesting to note that several of the exit-vector pairs share the same geometry, and thus the same hash.

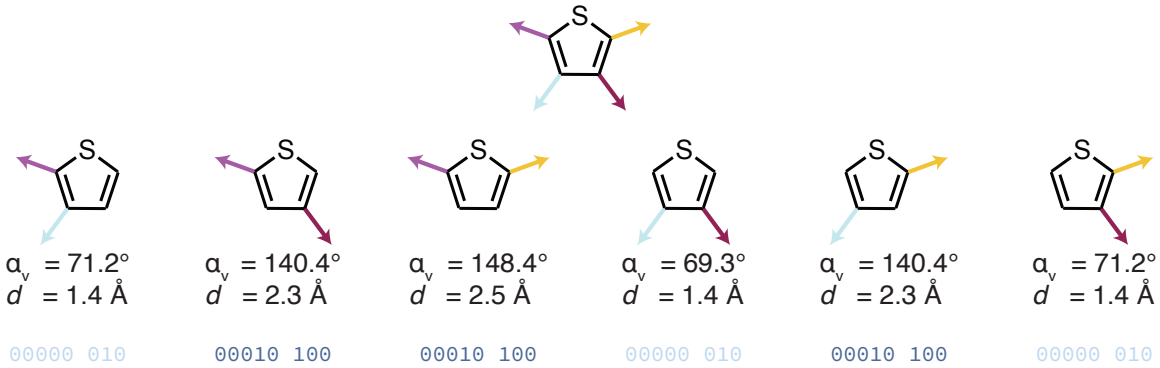


Figure 4.15. Calculating hashes for all the vector-pairs in each molecule in the expanded database, illustrated here for thiophene.

The final structure of the expanded database is a series of nested dictionaries, with the outermost keys as RegIDs for each heterocycle, then each of the nested internal dictionaries encapsulating information about the heterocycle necessary for searching, including the hashes and the exit-vector atom IDs corresponding to that hash. This structure is illustrated graphically in Figure 4.16. Stored under each RegID is the SMILES string for the heterocycle, the number of exit-vectors in that heterocyle, and a dictionary of the hashes corresponding to each pair of exit-vectors. Under each hash is stored the distance d between base atoms, the atom IDs of the vectors, and the angles α_{v}, α_1, and α_2 as defined in Figure 4.12. A subsequent dictionary was also created and distributed with the expanded dataset where the keys are the hashes, and the values are a list of the RegIDs of each molecule with at least one exit-vector pair corresponding to that hash.

Searching and aligning a heterocycle with two user-specified exit-vectors then becomes similar to that for a single vector, with the added step of hash indexing. Initially the hash is calculated for the user-specified exit-vector pair's geometry, and then all MoBiVic hetero-

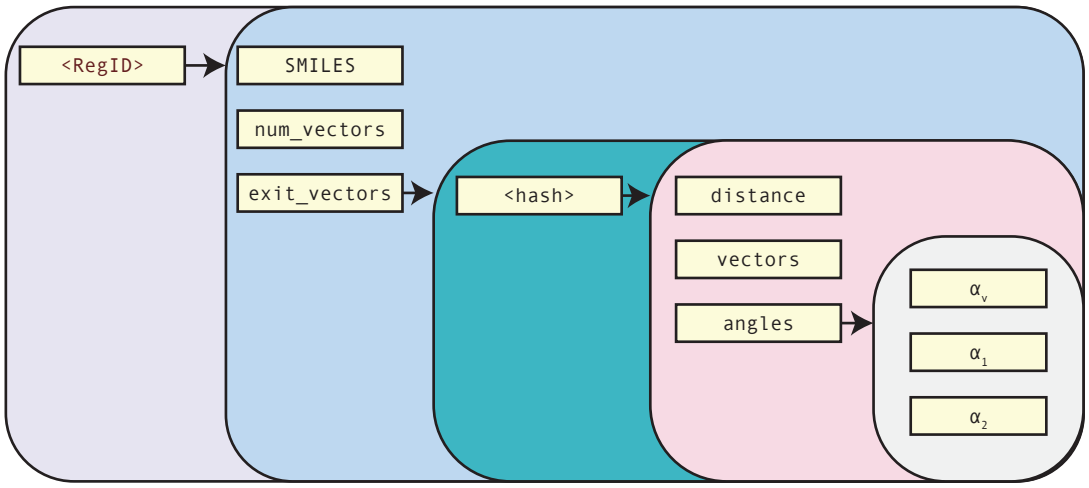


Figure 4.16. The structure of the database. Each rectangle represents a dictionary, with the yellow text indicating the keys. Keys with a following arrow indicate that their values are also dictionaries. Keys in angle brackets represent placeholders for the actual value of the property indicated.

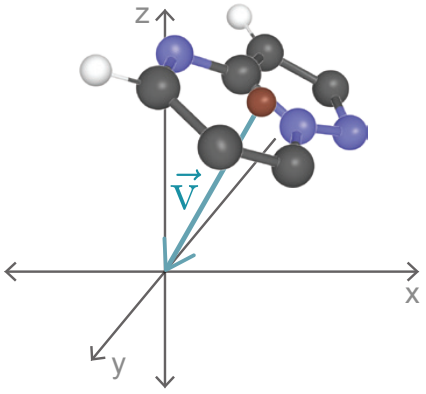
cycles whose database entries contain one or more vector pairs with that hash are retrieved from the hash-keyed dictionary.

After the hash-matching and retrieval, the alignment is similar to that of the one-vector case. First the centroid of the query and probe molecules (the non-weighted average of the heavy-atom coordinates) is translated to the origin (see Figure 4.17a). The atom IDs of the exit-vector pair(s) corresponding to this hash are retrieved from the database dictionary (see Figure 4.16) and these are used to construct the **P** matrix for the probe and the **Q** matrix for the query (as in Figure 4.17b). The Kabsch algorithm (as described in Section 4.2.1.6) is then used to calculate the rotation that minimises the RMSD of the probe onto the query, with the 4-row matrices now ensuring that the vectors are aligned onto each other. The probe is then rotated by this matrix to align it with the query, and this alignment scored for shape and ESP similarity as above. To account for asymmetry, the ordering of the vector

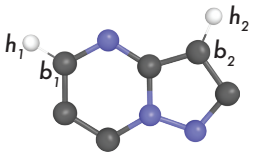
pairs is then switched in the $\mathbf{P}$ matrix to form the $\mathbf{P}'$ matrix as below

$$\mathbf{P}' = \begin{bmatrix} \longleftarrow & \vec{\mathbf{b}}_2 & \longrightarrow \\ \longleftarrow & \vec{\mathbf{h}}_2 & \longrightarrow \\ \longleftarrow & \vec{\mathbf{b}}_1 & \longrightarrow \\ \longleftarrow & \vec{\mathbf{h}}_2 & \longrightarrow \end{bmatrix}$$

This is then aligned to the unchanged $\mathbf{Q}$ matrix in the same manner. This has the effect of flipping the probe exit-vectors in the alignment. The new alignment is then scored for shape and electrostatic similarity, and the highest-scoring alignment of the two is returned, along with the coordinates of best alignment. To ensure that the returned SMILES string indicates the highest-scoring alignment, the `molAtomMapNumber` properties in RDKit are set for the highest-scoring alignment. This property determines which exit-vector is labelled as `[*:1]` or `[*:2]` in the output SMILES string for each probe ligand, and is set such that all the probe vectors that align with the vector labelled as `1` in the query molecule are also labelled as `1`. This enables the rapid visualisation of the highest-scoring alignment without needing to store many atomic coordinates, and allows for downstream computational functionalising.



(a) The vector ($\vec{v}$) defining the translation of the centroid to the origin. The red sphere indicates the calculated centroid of the molecule.



$$\mathbf{P} = \begin{bmatrix} \longleftarrow & \vec{\mathbf{b}}_1 & \longrightarrow \\ \longleftarrow & \vec{\mathbf{h}}_1 & \longrightarrow \\ \longleftarrow & \vec{\mathbf{b}}_2 & \longrightarrow \\ \longleftarrow & \vec{\mathbf{h}}_2 & \longrightarrow \end{bmatrix}$$

(b) The construction of the $\mathbf{P}$ matrix for exit-vectors of pyrazolopyridine shown. $\mathbf{b}$ defines the base atom and $\mathbf{h}$ the head (hydrogen) atom of the exit-vector pair. Their coordinates make up the rows of the $\mathbf{P}$ matrix, as shown.

Figure 4.17. Setting up the Kabsch alignment for a two-vector alignment.

An illustration of the full workflow for a two user-vector search is shown in Figure 4.18. The two-vector molecule is specified by the user as a SMILES string, for which an RDKit molecule is created, and the atom IDs of the atoms constituting the two exit-vectors determined and stored. After a 3D geometry is embedded and optimised, the hash is calculated for the specified exit-vectors. All molecules in MoBiVic with at least one pair of exit-vectors categorised by the same hash are retrieved, and these are each systematically aligned and scored against the query molecule in both possible orientations. For each probe molecule with the correct hash, the alignment with the highest total score is returned. These are then collated and returned to the user in order of highest total score, with the alignments and the scores available to the user.

4.2.1.8 Parallelisation

The search algorithm described above is computationally intensive, as it involves aligning and scoring against a large database of probe molecules. Recognising that aligning and scoring each molecule against the query is an easily parallelisable process, Python's standard `multiprocessing` library was utilised to increase the efficiency of the searching, allowing multiple probes to be aligned and scored simultaneously across available central processing unit (CPU) cores.

The alignment and scoring of probe molecules is thus distributed across the available CPU cores. For both one-vector and two-vector searches the total number of probe molecules being compared to the query is divided into near-equal sized batches, and each batch is allocated via a task queue to an available CPU. For a one-vector search, the entire MoBiVic library is divided into batches of 15 000, and for a two-vector search all the molecules retrieved in the hash indexing are divided into $n_{\mathrm{CPU}} + 1$ batches, where n_{CPU} is the number of available

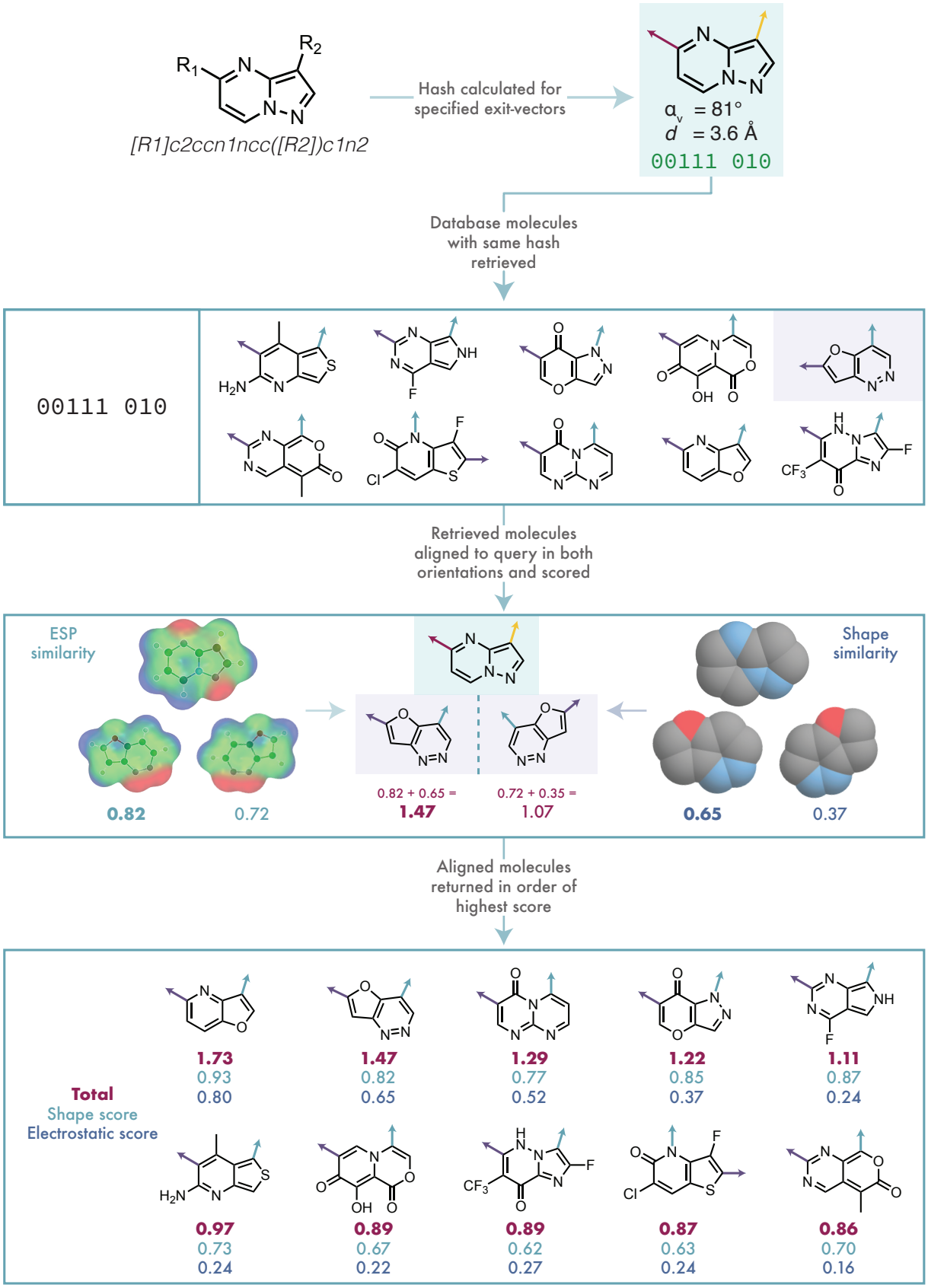


Figure 4.18. An illustration of the workflow for a two user-vector alignment.

CPUs. As the MoBiVic library of molecules is reasonably large, and recognising that each process requires read-only access to its contents (thus eliminating any risk of race conditions or data corruption), the dictionary of molecular properties (as described in Figure 4.16) was stored in shared memory. This is significantly more memory efficient as it prevents each process requiring its own local copy of the large database ($\approx$ 250 MB). An illustration of the parallelisation process is shown in Figure 4.19, using four CPUs as an example.

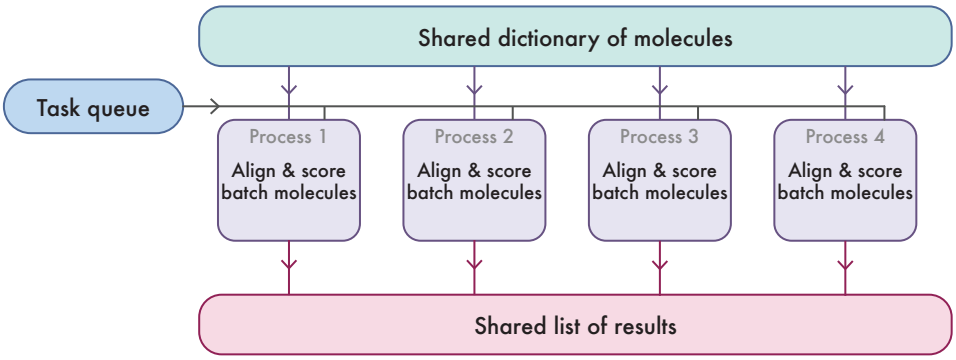


Figure 4.19. An illustration of the parallelisation process, spread across four processes (CPUs).

The task queue dynamically distributes the batches of molecules (passed as lists of RegIDs) to each of the processes, which then retrieves the necessary molecular information for each of the probes from the shared dictionary and executes the aligning and scoring logic. The tasks are distributed to the CPUs in an unordered manner to avoid the straggler problem, meaning that each CPU picks up a new batch as soon as it has finished the previous one, rather than idling until slower tasks are complete. After each batch has been aligned and scored, the highest-scoring alignment and the respective scores for each are passed to a shared list of results, which is then sorted by total score and returned to the user.

A flowchart outlining the overall alignment procedure is shown in Figure 4.20.

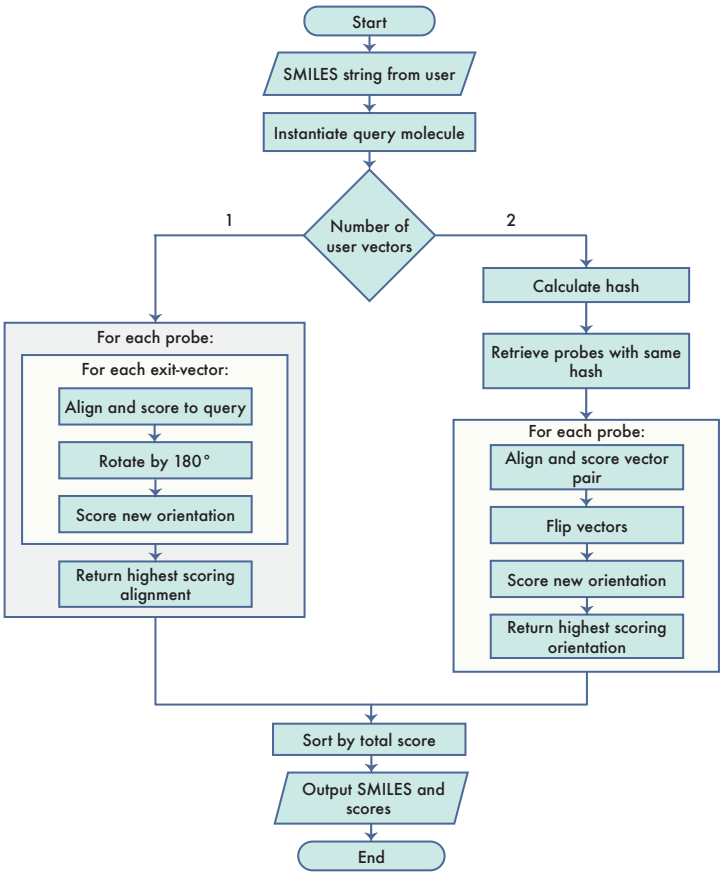


Figure 4.20. Flowchart describing the overall alignment and scoring process for a HCIE search.

4.2.2 Package Structure

The Python package `HCIE` - *the Heterocycle Isostere Explorer* was developed to implement the alignment algorithm and the large-scale library searching of query ligands described above to retrieve new bioisosteric pairings. The package was built on a minimum number of dependencies, and is designed to be efficient and easy to use. Once installed, a search can be carried out using two lines of Python code:

```
from hcie import VehicleSearch

search = VehicleSearch(smiles="[R]c1ccccn1", name="2-pyridine")
search.search()
```

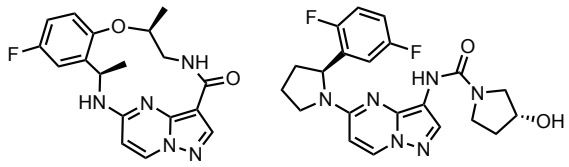
This then runs a search as described above, making use of the number of CPUs available on the local machine to efficiently parallelise the alignment process. For the search above a directory entitled `2-pyridine_hcie_results` is created, and a `.txt` file with results for each screened molecule, ordered by total score is deposited there. Also included in this directory is an `.sdf` file containing the alignments of the 50 highest-scoring molecules, and a `.png` file of their molecular structures. The parameter determining how many molecules are printed to the `.sdf` and `.png` files is user-adjustable. For a two-vector search, both vectors are specified with `[R]` in the SMILES, and the vectors are numbered in the output automatically.

4.3 Results and Discussion

4.3.1 Bioisosteres for Pyrazolopyridine

To evaluate the results of searching this database with a two user-vector query, 3,5-disubstituted pyrazolopyridine was selected as an example. 3,5-Disubstituted pyrazolopyridine appears as a structural motif in the recent FDA-approved *ROS1* tyrosine kinase inhibitor (ROS1TKI) repotrectinib (marketed as Augtyro™), showing anti-tumour activity in non-small cell lung cancers, including those with mutations rendering them resistant to early-generation therapies.[311] This pyrazolopyridine originally featured in the first-in-class tropomysin kinase inhibitor larotrectinib (approved by the FDA in 2018, and marketed as Vitrakvi™), where the pyrazolo-nitrogen is suggested to form a hydrogen bond with Met-592 in the hinge binding region, and is thus important for both potency and selectivity.[89,312] This is an intriguing heterocycle due to the number and placement of the nitrogen heteroatoms within the rings, including a nitrogen at the ring-junction, which are likely to give it an ESP that is difficult to predict. Manually identifying electrostatically similar bioisosteres is there-

fore non-trivial. Furthermore, at the time of approval the price of larotrectinib (as reported in a 2018 Form 8-K filing with the United States Securities and Exchange Commission) was USD 32 800 a month, and repotrectinib marketed at USD 30 740 a month.[313,314] These high prices render these molecules as likely candidates for the development of cheaper 'me-too' drugs, and thus bioisosteres of the key hinge-binding motif might be of use industrially.



Repotrectinib **Larotrectinib**

To find proposed bioisosteres of this motif a HCIE search was performed (as illustrated in Figure 4.18). The computed hash was `00111010`, and 101 294 probe molecules (nearly 20% of MoBiVic) with the same hash (and thus similar exit-vector geometry) were identified. Each of these was aligned and scored to the query, taking 122 seconds in total, and the results of this search are illustrated in Figure 4.21.

Pleasingly the 3,5-disubstituted query was returned from the database with a perfect shape and ESP score, suggesting that the alignment and scoring algorithm employed here is reliable. Furthermore, all the top returned probes in Figure 4.21 panel A appear to be chemically reasonable, suggesting that the filters outlined in Section 3.2 have succeeded in removing many of the less realistic molecules.

Inspecting Figure 4.21 panel A, all but **4.6**, **4.7**, and **4.9** have an H-bond acceptor in the correct position to form the crucial Met-592 interaction, and all of them appear to have an exit-vector geometry corresponding to that of the 3,5-disubstituted pyrazolopyridine input. This suggests that the hash-based searching algorithm is successful in retrieving molecules

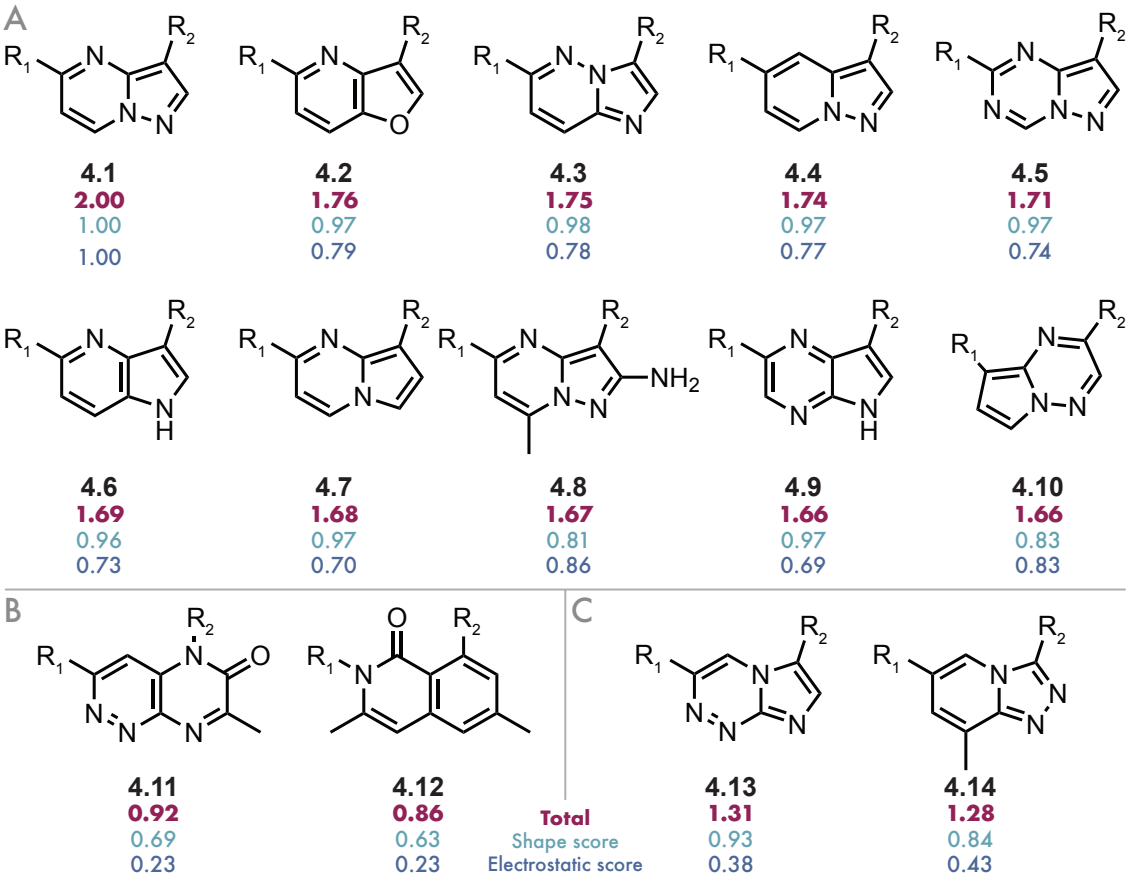


Figure 4.21. **A** | The top 10 returned molecules from a HCIE search of 3,5-disubstituted pyrazolopyridine. **B** | Examples of low-scoring molecules returned from the same search. **C** | Examples of visually plausible bioisosteric candidates that are nonetheless scored poorly by this methodology.

with a similar arrangement of vectors, and the short search time (just over two minutes) highlights its efficiency.

It is interesting to note that whilst the shape scores are universally high across these results (> 0.8), the ESP scores are lower and more varied. That the shape scores are high is not surprising; all of the returned scaffolds are 6,5-fused heterocycles and all bar **4.10** are aligned such that rings of the same size always overlap in space. The lower shape similarity score of **4.8** can be attributed to the inclusion of the 2-amino and 7-methyl substituents, which significantly increase the steric bulk about the central core when compared to the query. The lower shape similarity score of **4.10** is due to the returned alignment, with its 5-membered

ring exit-vector aligned to that of the 6-membered ring in the query, thus resulting in a poorer shape similarity than if the exit-vectors had been aligned to the query in the opposite manner. This demonstrates the balance to be struck between shape and ESP similarity when considering bioisosteres for scaffold-hopping; the returned alignment of **4.10** has the highest ESP score of all returned non-query heterocycles except **4.8**, for which the heterocyclic scaffold is the same as the query only decorated with electron-donating substituents. As the weightings of shape and ESP contributions to the total score were equal for this search, the increased ESP similarity that arose from the better aligning of the nitrogen-nitrogen bond in **4.10** with that of the query was significant enough to outweigh the penalty caused by the poorer spatial overlap. By inspection alone this is not immediately obvious, and it is likely that **4.10** in the alignment proposed here would be overlooked in a traditional, manually-designed scaffold hop.

The lower ESP scores can be explained by the sensitivity of the ESP to the precise arrangement of heteroatoms with the aromatic ring system.[81] Different elements have different electronegativities, and so their precise arrangement within the ring has a significant effect on the overall electronic distribution and thus the dipole moment. When comparing **4.6** and **4.9** to **4.1**, it is clear that introducing a 'pyrrole-like' nitrogen (thus an H-bond donor) in place of the pyrazole's 'pyridine-like' nitrogen (an H-bond acceptor) results in a significant difference in ESP similarity. Replacement of that nitrogen with a carbon, as in **4.7**, also has an impact on this similarity for the reasons outlined above. It is interesting to note that the introduction of substituents has less of an effect on the ESP similarity scores than changing the constitution of the heteroatoms; **4.8** shares the same cyclic atoms as **4.1** and has the highest ESP similarity score.

Figure 4.21 panel B shows two of the lowest scoring heterocycles from the search results. Both **4.11** and **4.12** are 6,6-fused heterocycles, and thus it is not surprising that their shape scores are comparatively lower than those in panel A. This demonstrates that the shape scoring method used in this implementation is more discerning of these subtle differences than that outlined in Chapter 2, where rings of different sizes to the query commonly scored highly. Visually it is clear that these heterocycles bear little resemblance to the pyrazolopyridine query molecule, and thus it is reassuring that they are scored significantly lower than the molecules in panel A.

Illustrated in Figure 4.21 panel C are two heterocycles which visually seem plausible as bioisosteric candidates, and could reasonably be expected to be included in a manually compiled list of candidate molecules in a medicinal chemistry campaign. It is interesting to note that these seemingly plausible heterocycles score poorly out of the 5,6-fused cycles returned by the search algorithm, suggesting that they might be less adequate bioisosteres. Both **4.13** and **4.14** score highly on shape similarity but the electrostatic similarity scores for both are much lower. This subtle distinction would be difficult to discern by eye, and as such these examples highlight the value of the HCIE methodology for de-prioritising candidates that might otherwise advance in a human-led campaign.

Inspection of the 3D geometries extracted from the `.sdf` file coordinates, as displayed in Figure 4.22, demonstrate that the alignments are as expected and correspond to the vector arrangements given in the SMILES strings.

To investigate whether any of these proposed heterocycles have previously been used as bioisosteres in scaffold-hopping studies, the 3,5-pyrazolopyridine query was searched in the SwissBioisostere database.[220] This retrieves scaffolds from the ChEMBL database that have

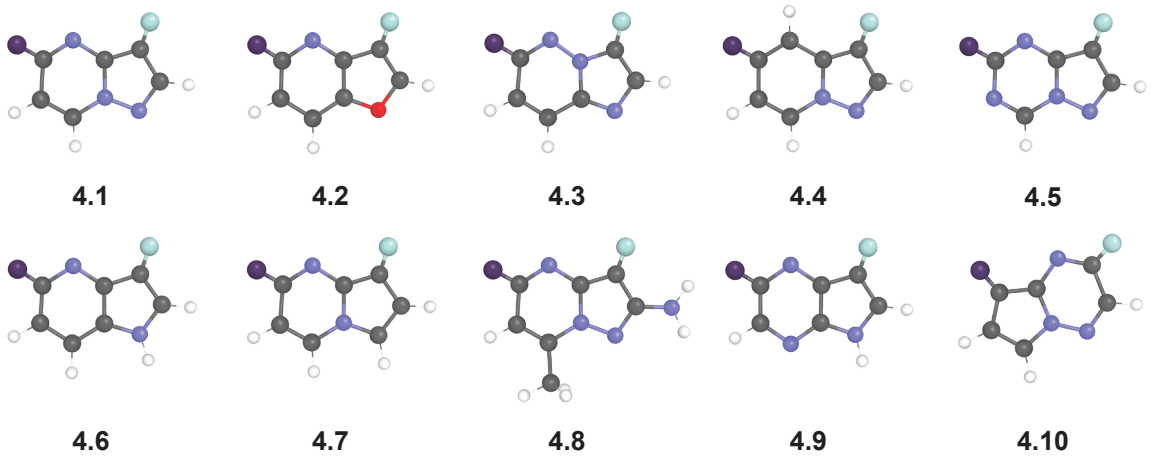


Figure 4.22. The 3D alignments of the top 10 highest-scoring molecules returned in the HCIE search for 3,5-disubstituted pyrazolopyridine. The exit-vectors are coloured in purple and pale green.

been substituted for 3,5-pyrazolopyridine in the literature, recording the number of times that particular substitution has been made within ChEMBL and the proportion of these substitutions that improves, retains, or reduces reported bioactivity. The SwissBioisostere search returned 44 unique aromatic heterocyclic scaffolds. Of these, 14 were heterocycles with substituents not included within MoBiVic or substitution patterns outside of the searching methodology (for example *N*-substitution on an amino group, or *O*-substitution on a phenol group) and so would never appear in HCIE results. A further seven of these did not have the same exit-vector geometry pattern as that of the query, and so would have a different hash as calculated by this methodology and thus also would not appear in HCIE results.

Pleasingly **4.3** was the bioisostere with the highest number of recorded replacements in ChEMBL, and had the best bioactivity improvement. Of the 33 MMPs existing in ChEMBL, 30 of these improved or retained bioactivity. **4.4** appeared as the 6th most frequent scaffold hop in the SwissBioisostere results with 5 MMPs, all of which retained or improved bioactivity. It is not surprising that **4.6** or **4.9** do not appear in the returned results as these lack an H-bond acceptor in the 'pyrazole-like' nitrogen position as aforementioned. However, **4.2**,

4.4, **4.5**, **4.7**, **4.8**, and **4.10** have not been used as a bioisosteric scaffold for 3,5-disubstituted pyrazolopyridine in the reported literature. Reaxys and Scifinder searches of the published literature revealed that each of these ring systems has previously been synthesised, thus suggesting a lack of synthetic accessibility is unlikely to be the reason for these pairings never having previously been attempted. This does not mean, however, that these are not effective bioisosteres of 3,5-pyrazolopyridine, just that they have yet to be characterised.

That the searching method described here is able to retrieve known and effective bioisosteric pairings validates the robustness of the vector-based alignment algorithm, and the hash-based categorising and searching methodology. The absence of many of the top 10 proposed bioisosteres from the literature, despite their proven synthetic accessibility, illustrates the potential of exploration of aromatic heterocyclic chemical space for novel bioiososteric pairings.

4.3.2 Rationalising the Activity of Inhibitors of the NLRP3 Inflammasome

In recent years, there has been a significant interest in the role of inflammation in neurodegenerative disorders such as Alzheimer's and Parkinson's disease.[315–318] Microglial activation (one of the key enactors of the innate immune response in the central nervous system (CNS)) has been shown to be associated with the characteristic amyloid-beta (Aβ) plaques found in the brains of Alzheimer's disease (AD) patients.[319] The NLRP3 (NOD-, LRR-, and pyrin-domain containing protein 3) inflammasome assembles in activated microglia, and its assembly leads to increased cleavage (and thus activation) of caspase B, and downstream release of the cytokine interleukin-1β.[320] Aβ is known to activate the NLRP3 inflammasome, and

Ising *et al.* showed in 2019 that its activation drives the tau pathologies linked to cognitive decline in AD patients.[321–323] Difficulties in developing clinically-effective drugs targetting Aβ, coupled with an interest in targeting downstream pathways of Aβ deposition have lead to a recent surge in efforts to discover novel small-molecule inhibitors of this key aspect of the innate immune response for the treatment of neurodegenerative disorders and peripheral inflammation.[324–328]

The Brennan Group's long-standing interest in small-molecule therapeutics for the treatment of neurodegeneration, coupled with the large number of biologically-characterised NLRP3 inhibitors available in the patent and published literature, prompted an investigation into whether the bioactivity of these inhibitors could be rationalised in terms of shape and electrostatic similarity.[329] A dataset of 8974 unique small-molecule inhibitors of the NLRP3 inflammasome, extracted from patent and published literature and annotated with IC_{50} data, was purchased from the commercial provider GOSTAR®, and provided by collaborators at Exscientia. In order to compare fairly the bioactivities of these data to the shape and ESP similarity scores, matched molecular series (MMS) needed to be extracted from the dataset. These are groups of molecules where the only difference between each molecule in a group is a single, well-defined molecular substitution; for the molecules relevant to investigation this substitution would be aromatic heterocycles within MoBiVic.

To identify these MMS the `mmpdb` Python package described by Dalke *et al.* was used to group molecules.[330] This represents an efficient, SQLite-backed implementation of the Hussain-Rea fragmentation algorithm, and is fast at fragmenting large datasets into MMPs and MMS. This algorithm, first described in 2010, identifies MMPs by sequentially cutting acyclic bonds in a molecule according to a series of pre-defined rules.[223] The fragments that result are canonicalised and stored in a dictionary structure with the constant fragment (that

which is the same between pairings) as the key, and the variable fragments (those parts of the molecule that differ in the pair) as values. This is repeated for all molecules in the dataset, and then all the variable fragments (values) associated with a particular constant fragment (or key) define a MMS.

For the purposes of this investigation, the rules defining which bonds could be cut to generate the fragments were adapted such that only MMS where the variable fragments were aromatic heterocycles present in MoBiVic were generated. The SMARTS-based rules for bond-cutting were defined as follows:

1. Only bonds from aromatic atoms to non-aromatic atoms are cut.

2. Bonds from aromatic atoms to substituents present in MoBiVic are not cut.

3. Double and triple bonds are not cut.

The following SMARTS expression encapsulates the rules above, and was used to define the cutting pattern for this analysis:

```
'[R]!@!=!#[!#0;!#1;!#9;!#17;!$([CH3]);!$([OH]);!$([NH2]);!$([CF3]);!$([OCH3])]'
```

Following the fragmentation and indexing, MMS containing fewer than ten molecules were discarded and the remaining series were each ordered by IC_{50}. The variable fragment of the most potent ligand in each series was taken as the query, and all other variable fragments from the same series were aligned to the relevant exit-vectors in the query and scored using the HCIE methodology. The total score was initially calculated as an equally weighted sum of the shape and ESP similarity components. These datasets were then divided into those representing single-vector variable fragments and two-vector variable fragments. For each of

these groups the series with the highest number of ligands were selected, having removed any series where the bioactivity data was too discretised to extract meaningful statistical results (an example of this is shown in Figure 4.23). These horizontal groupings are due to a number of ligands in each series being assigned the same bioactivity value in their data source, and is likely caused by saturation in their respective assays.

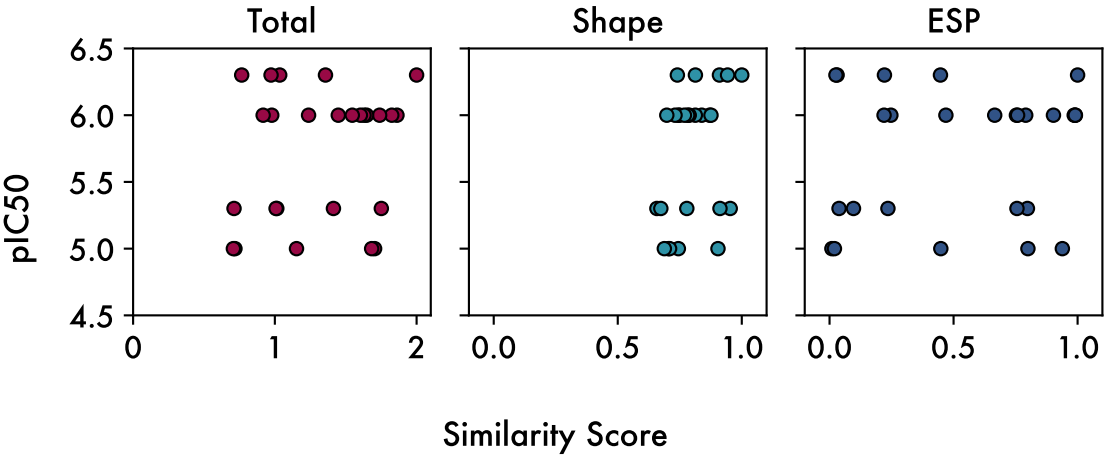


Figure 4.23. An example of a matched molecular series where the bioactivity data is too clustered to extract meaningful data.

When these statistically difficult series had been removed, five series remained for analysis. Three of these represented single-vector substitution series, and two of these two-vector substitutions. The constant fragments defining these series, and the number of unique ligands in each series, are shown in Figure 4.24.

Series D is derived from **CRID3** (also known as MCC950), a commercially available and potent NLRP3 inhibitor with an pIC_{50} of 8.09 in human monocyte-derived macrophages.[331] Although displaying high potency *in vitro*, trials of **CRID3** in humans were stopped due to concerns about hepatotoxicity, and the compound displayed poor brain penetrance hindering its use in AD therapies.[332]

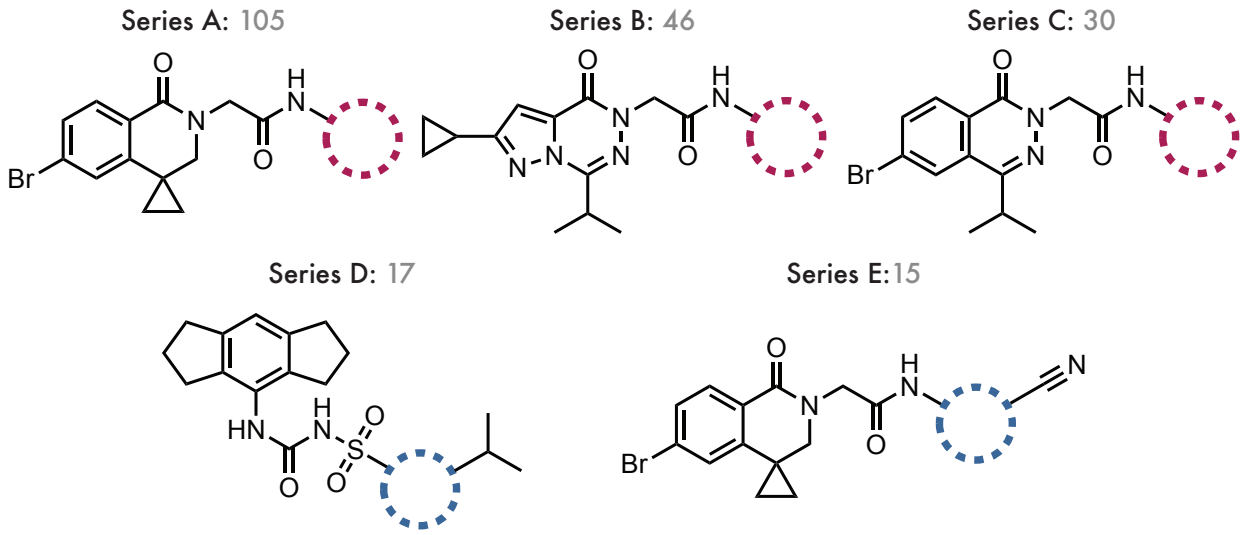
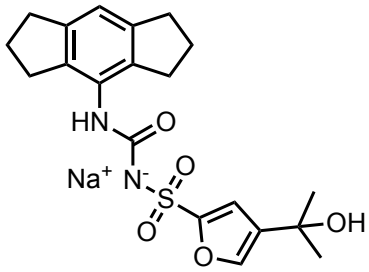


Figure 4.24. The constant fragments defining the five matched molecular series used in this analysis, and the number of unique ligands present in each series.



CRID3

Series E is a subset of series A, where both the original series A constant fragment and an aromatic nitrile substituent are held constant across the series. Although it would have been preferable to have more ligands in each of the two-vector series D and E, previous difficulties in obtaining datasets for these two-vector 'scaffold-hops' suggest that these datasets are the largest available.

Initially the equally-weighted total score was correlated against the pIC_{50} values for each series, and the plots of these (along with the Pearson correlation coefficient values for each series) are shown in Figure 4.25, and the full breakdown of the calculated correlation statistics in Table 4.3. The Pearson correlations between equally-weighted total score and bioactivity

gave positive scores in the range of 0.43 - 0.65, and all with a probability value below 0.05. As these scores are a much-simplified representation of the complex factors that determine binding affinity (for example entropic contributions, macro-conformational effects, or solvent interactions are not directly considered), a very high correlation coefficient would not be expected. Furthermore, correlation with experimentally derived binding affinities will always include a degree of experimental variability, which further reduces the expectation of perfect correlation. These results demonstrate that there is a correlation between the shape and ESP similarities of aromatic heterocycles in these series and their binding affinity, thus suggesting that a high combined shape and ESP score is a useful metric for proposing potential new bioisosteres.

Table 4.3. The correlation statistics for the equally-weighted total scores for the NLRP3 MMS.

Series	Pearson Coefficient	p-value
A	0.43	5.5×10^{-6}
B	0.58	2.7×10^{-5}
C	0.58	7.7×10^{-4}
D	0.62	7.8×10^{-3}
E	0.65	8.2×10^{-3}

As with the results described in Section 4.3.1, all series except E show a general trend for shape scores being higher and with a more compact distribution than those of the ESP scores. As the methodology for scoring is the same in both cases, it is likely that the reasons suggested above hold for these NLRP3 inhibitor data.

Inspection of the distributions of the equally-weighted total score and the constituent scores for each series, combined with the differences between shape and ESP similarities described for the bioisosteres of 3,5-pyrazolopyridine above, suggested differing contributions to the overall correlation between the two metrics which might not be best represented in an equal

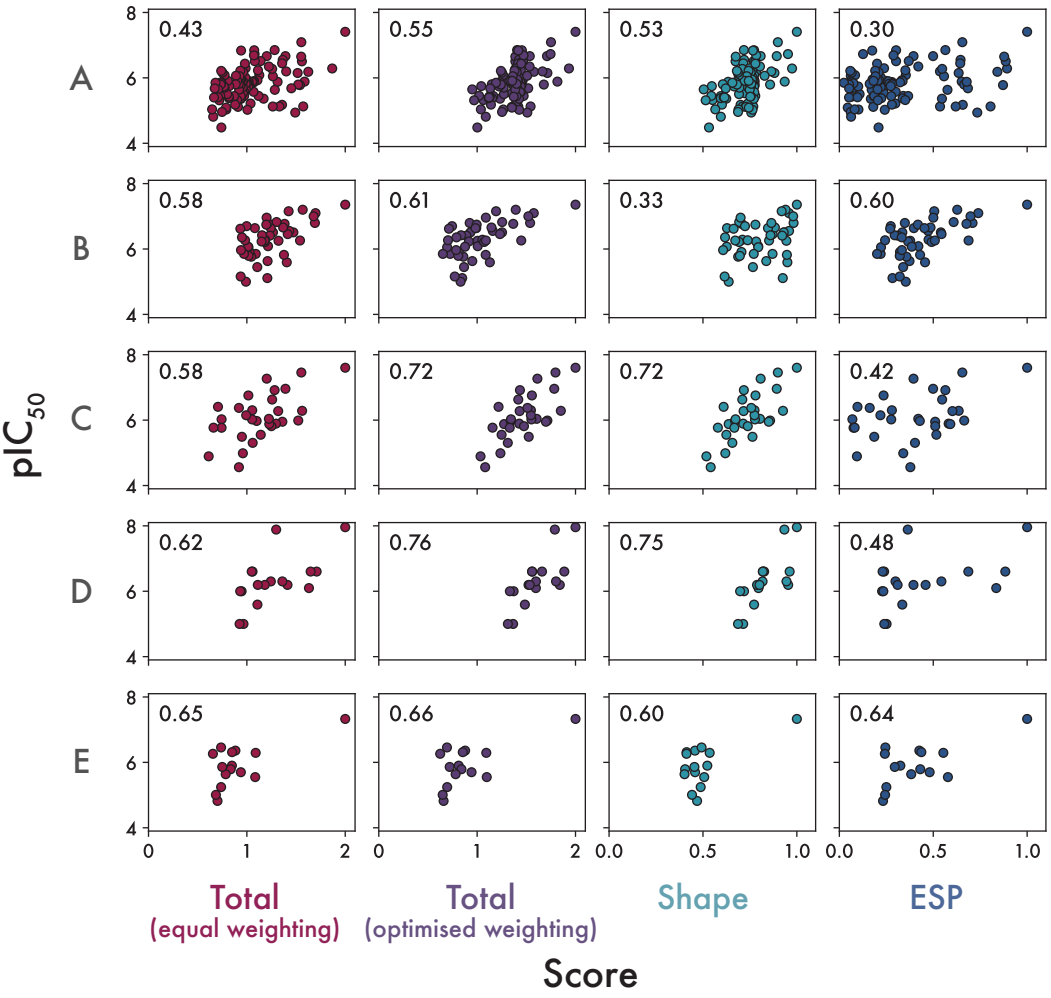


Figure 4.25. The correlation plots of the NLRP3 inhibitor matched molecular series, for both equally-weighted and optimised total score. The Pearson correlation coefficient scores are displayed for each plot.

weighting. For example, in series A the shape scores correlate to binding affinity with a higher coefficient than that of the total score. To identify whether an unequal weighting in the total score would lead to better overall correlation with the binding affinity, a constrained optimisation approach using the sequential least squares programming (SLSQP) algorithm as implemented in SciPy was employed. The objective function was defined to minimise the negative Pearson correlation coefficient between the pIC_{50} and the weighted sum of the ESP and shape scores, with each weight constrained to the interval [0, 1], and the sum of weights

equalling 1. The final weightings which lead to the most positive correlation coefficient for each series are shown in Table 4.4, and are also displayed in Figure 4.25.

Table 4.4. Optimised weightings, as calculated using the SLSQP optimisation algorithm for each series, and the Pearson correlation coefficients for the total scores calculated using the optimised weightings.

Series	Weighting		Pearson	
	Shape	ESP	Coefficient	p-value
A	0.90	0.10	0.55	1.6×10^{-9}
B	0.22	0.78	0.61	1.2×10^{-5}
C	1.00	0.00	0.72	6.3×10^{-6}
D	0.93	0.07	0.76	4.6×10^{-4}
E	0.41	0.59	0.66	1.0×10^{-2}

As would be expected from inspection of Figure 4.25, many of these weightings deviate from the initial equally-weighted starting values. Correlating the total scores calculated with optimised weightings against the pIC_{50} data leads to improvements in the Pearson correlation coefficients across all series, with these now lying in the range 0.55-0.76. Interestingly there appears to be no consistent weighting across the series, with series A, C, and D showing a skewing towards shape similarity whereas B and E are skewed towards electrostatic similarity. Indeed for series C the optimised weighting takes no account of the ESP similarity scores, with the best correlated total score mirroring exactly the shape similarity scores. These results suggest that the relative importance of steric and electrostatic complementarity in determining bioisosteric suitability is target and ligand dependent, with different series behaving as more 'shape-like' or more 'ESP-like'. A retrospective optimisation analysis such as this one on existing data in a ligand discovery project would enable the determination of the optimum weightings for the project in question, and these weights could then be used for prospective molecule design.

This analysis rationalises the pIC_{50} of the ligands in each series of NLRP3 inhibitors by assessing their similarity to the most potent in each series. This shows that although some series are better represented by their ligands' shape or ESP similarity, there is no clear pattern to which will be a better representation. If prior data is available, a retrospective optimisation analysis such as that described here is likely to lead to better predictions of bioactive bioisosteres than a simple equal weighting. However, the correlations between equally-weighted total scores and pIC_{50} (as shown in Figure 4.25 and Table 4.3) are sufficiently robust to justify using equal-weighting as a reasonable initial approach in the absence of prior data.

Furthermore, this analysis was conducted using only ligands already synthesised and assayed as part of the discovery campaigns for each series, leaving open the possibility that more potent ligands yet to be designed, synthesised, or assayed exist in MoBiVic. At the time of writing the biological characterisation of these bicyclic NLRP3 small-molecule inhibitors is mostly restricted to their cellular bioactivity, thus it is likely that further optimisation of their molecular structure will need to be made as more ADMET profiling is performed.[328] Heterocyclic replacement is a well-known strategy for optimising pharmacokinetic and physicochemical properties in lead-like molecules.[107] As the relationship between shape and electrostatic similarity (as calculated using the methodology described here) and potency has been demonstrated, there is scope for the HCIE methodology to aid in the design of analogues that retain bioactivity while offering alternative physicochemical or pharmacokinetic properties, thus expanding the chemical space for discovering superior bioisosteres. Following this validation of HCIE's ability to reproduce the structure-activity relationships across these inhibitor series, the subsequent analysis benchmarks the capacity of the software to identify novel bioisosteres of a widespread heterocycle in medicinal chemistry.

4.3.3 A Novel Class of Bioisosteres for 2-Pyridine

A 2024 analysis by Marshall *et al.* found that the proportion of FDA-approved drugs containing a nitrogen heterocycle increased from 59% in the period spanning 1938-2012 to 82% between 2013-2024.[79] In the latter period, pyridine supplanted piperidine as the heterocycle most frequently appearing in these small-molecule drugs, appearing in 54 of the 321 unique molecules approved in this period. The authors noted that 90% of these pyridines were substituted in the *ortho*(2)-position, thus making 2-pyridine the most frequently used aromatic heterocyclic moiety in approved drugs.

To investigate whether this methodology could propose interesting bioisosteres of this important motif in medicinal chemistry, 2-pyridine was searched through HCIE. The search took 12 minutes 10 seconds, and returned all 546 271 probe ligands aligned, scored, and ranked in order of highest to lowest, with 2-pyridine being returned as the top-ranked match (with a perfect score of 2.0). These results are shown in Figure 4.26. 2-thiophene was returned as the highest scoring proposed bioisostere (with an equally-weighted total score of 1.72). Thiophene is a known bioisostere of pyridine, and a search in SwissBioisostere revealed that the 2-pyridine to 2-thiophene substitution had been made 588 times in ChEMBL, improving or retaining bioactivity in 514 of these cases.

As alluded to previously, finding a dataset with a sufficient number of biologically characterised aromatic heterocyclic MMPs is challenging, therefore defining the 'ground-truths' for calculating enrichments to benchmark the HCIE methodology was not straightforward. In order to calculate enrichments for the HCIE methodology, and thus determine what proportion of the results are already known bioisosteres, the results of the SwissBioisostere search for 2-pyridine were used as the known active ligands. This is valid as these SwissBioisostere pairings are extracted from the literature and as such are biologically characterised. They

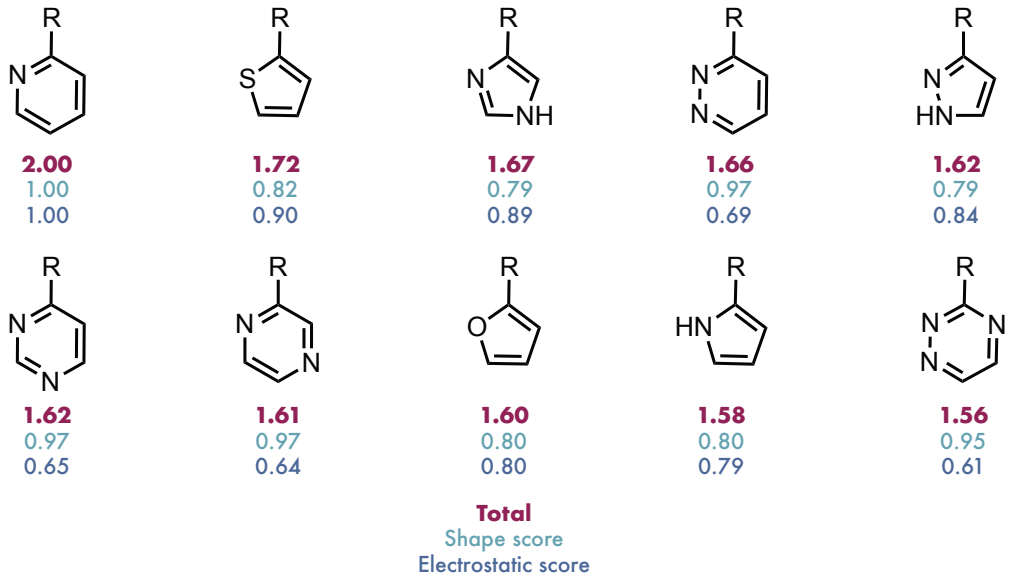


Figure 4.26. The top 10 returned results for 2-pyridine in the alignment of highest total score, and their respective scores.

are also annotated with the effect that the pairing has on the bioactivity, thus it is possible to include only those pairings that reliably retain or improve bioactivity. The SwissBioisostere search for 2-pyridine gave 7072 results. These were then filtered to remove all pairings for which the frequency of reduced bioactivity outnumbered the frequency of either retained or improved bioactivity, and further filtered to remove any molecules that were not present in MoBiVic. This left a dataset of 389 known, aromatic heterocyclic bioisosteric pairings for 2-pyridine, which were taken for this analysis as the known active ligands. The enrichment factors were then calculated from the HCIE search results using Equation 2.4, and are displayed in Table 4.5.

The enrichment factors for the top fractions of the HCIE results are large, showing that HCIE is able to retrieve known active bioisosteres significantly more effectively than a random draw. The large size of MoBiVic is likely to enlarge these early enrichment factors, as the probability of drawing nine active ligands at random from a database of over 500 000

Table 4.5. Enrichment factors calculated for 2-pyridine using both HCIE and ElectroShape. An enrichment factor of 1 indicates a performance no better than random, and a factor larger than one indicates a better than random ranking of ligands.

	HCIE		Electroshape	
Top N Molecules	Retrieved	EF	Retrieved	EF
10	9	1264	5	702
25	15	843	17	954
50	23	646	23	646
100	31	435	23	323
10%	350	9.0	301	7.7
25%	376	3.9	324	3.3
50%	388	2.0	364	1.9
75%	389	1.3	381	1.3
100%	389	1.0	389	1.0

molecules is very small, however the proportion of known ligands in the top 10, 25, and 50 molecules is very encouraging. The top 10 returned molecules from the HCIE search (excluding the 2-pyridine query) returned nine known bioisosteres from the SwissBioisostere results, with the top 25 returning 15, and the top 50 returning 23 known bioisosteres, all of which are known to improve or retain bioactivity. It is important to consider, when interpreting these enrichment factors, that a high-scoring molecule in the HCIE results that is not found in the known active ligands is not necessarily a false positive (as would be the case when comparing these to enrichments calculated with curated libraries such as the Directory of Useful Decoys).[333] As MoBiVic contains both molecules that have never been synthesised and molecules that have never been considered as bioisosteres, there exists the possibility that molecules in the HCIE results but not the SwissBioisostere known actives could be bioisosteres, but have yet to be tested or recognised as such.

In order to compare the HCIE results to those of a well-established and widely used virtual screening method, the above process was repeated using the Electroshape ultrafast shape recognition (USR) method (see Section 1.4.2.1).[250] Descriptors were calculated for

each molecule in the database using the Electroshape implementation in the Python Open Drug Discovery Toolkit, and their similarities to the pyridine descriptor calculated.[334] The Electroshape calculations and searching took 11 minutes, and the results (ranked in order of highest similarity) were used to calculate the enrichment factors as above, which are displayed in Table 4.5. The number of known bioisosteres retrieved is less than or similar to the HCIE results at all proportions except the top 25 molecules, for which the Electroshape search retrieved two additional bioisosteres. Pleasingly in the top 10 ranked ligands HCIE retrieves nine known bioisosteres to Electroshape's five, and in the top 100 ranked molecules HCIE identifies 31 whereas Electroshape only retrieves 23.

That HCIE is able to retrieve a comparable or better number of bioisosteres compared to Electroshape, a benchmarked and frequently-used tool in computational medicinal chemistry, demonstrates that the vector-based search is a valid methodology for searching MoBiVic. Crucially, the HCIE methodology returns the alignment of highest similarity, and highlights the exit-vector(s) corresponding to this alignment, thus providing useful structural information about the proposed bioisosteric pairings. These exit-vector designations are also useful for further computational compound design, as the output from HCIE can be passed directly into a virtual compound enumeration or docking pipeline without further intervention, whereas the USR results require manual inspection and exit-vector designation. That HCIE only took 70 seconds longer than Electroshape to search the entire MoBiVic library, whilst returning more information serves to highlight the efficiency of the algorithm.

To explore the possibility of discovering new bioisosteres of 2-pyridine within these results, the top 50 highest scoring moleclues that did not appear in the SwissBioisostere results were examined. Amongst these, it was interesting to observe that a number of unusual 5,5-bicyclic molecules scored highly for similarity to the 2-pyridine query, which are displayed in Fig-

ure 4.27. 5,5-bicyclic aromatic heterocycles are seldom encountered in medicinal chemistry, however examples of their inclusion in anti-microbial candidates and enzyme inhibitors exist in the literature, suggesting that this large class of heterocycles contains untapped biological potential.[335–337] These results suggest that the molecules shown in Figure 4.27 below could act as interesting new bioisosteres of 2-pyridine.

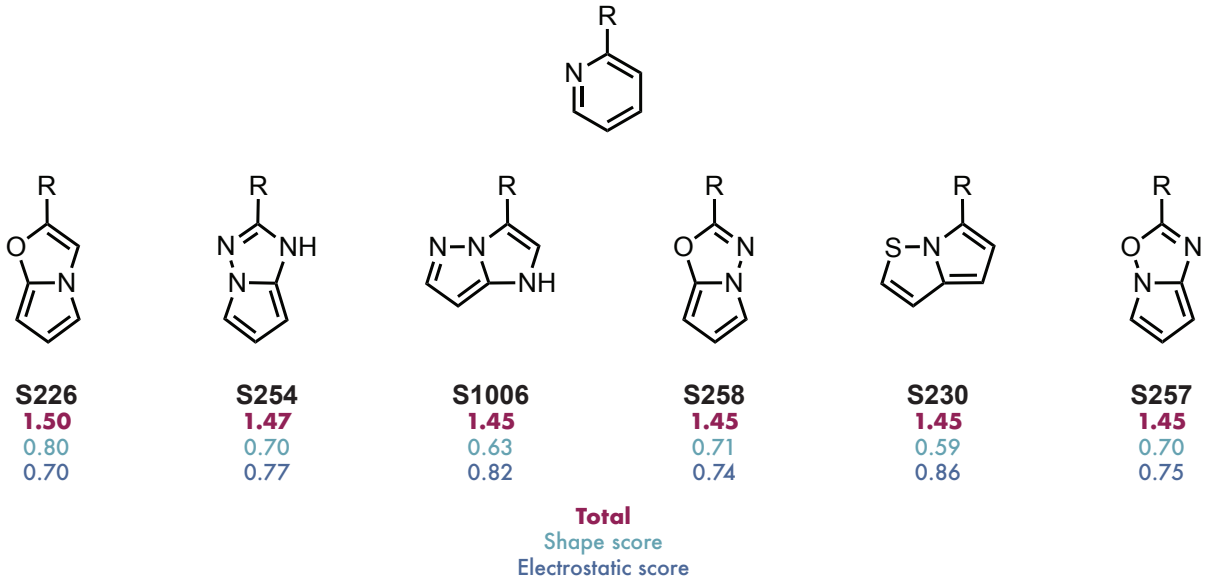


Figure 4.27. The proposed 5,5-bicyclic bioisosteres of 2-pyridine in the alignment of highest similarity, and their scores.

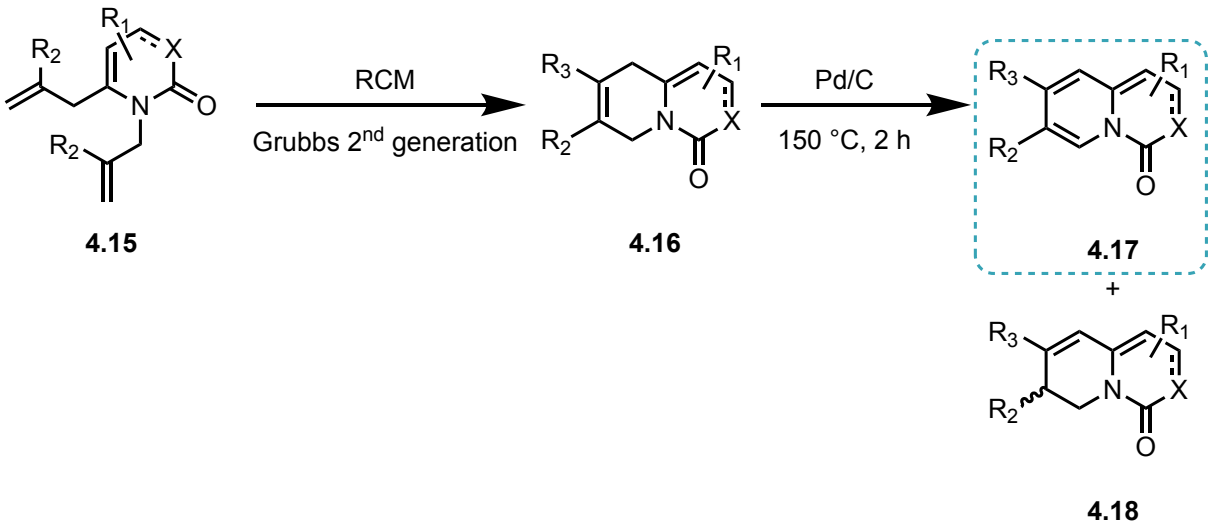
The alignments of all of these molecules, with the exception of **S1006** and **S230**, are unsurprising, with the long axis of the 5,5-bicycles aligned along the axis of the 2-pyridine exit-vector. Interestingly, **S1006** and **S230** are aligned with their long axes skewed relative to that defined by the 2-pyridine exit-vector. This allows the ring-junction nitrogen in both cases to better align with the pyridine nitrogen of the query, giving a larger ESP similarity value that outweighs in both cases the effect on shape similarity caused by the skewed alignment. Whether this is an acceptable trade-off in general will depend on the precise nature of the binding pocket in the target.

Of the molecules in Figure 4.27, only **S254** and **S1006** have previously been reported in the literature with the given substitution pattern, with **S254** appearing in the Markush structures of two separate patents and **S1006** having a synthetic route reported in 1999 (although its first inclusion in a bioactive molecule wasn't reported until 2024).[338–341] Until the beginning of 2024 there were no reported syntheses in the published literature of any of the remaining molecules. However, pleasingly in 2024 a molybdenum catalysed deoxygenative coupling strategy to access various heteroatomic scaffolds, including **S226**, was reported by Wang *et al.*[342] That syntheses of these molecules are continuing to be reported indicates that there is promise in developing routes towards as-yet unsynthesised 5,5-bicyclic systems, and their biological evaluation.

4.3.3.1 Synthesis

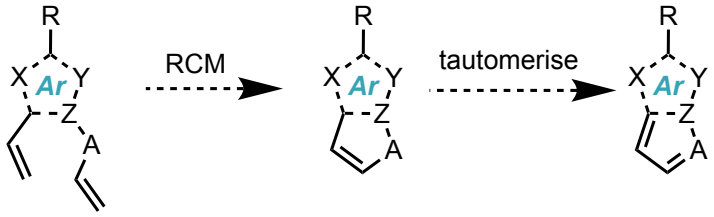
Despite there existing synthetic routes for two of the heterocycles in Figure 4.27, there exists no more general route to accessing these 5,5-heterocycles, which restricts their utility as bioisosteres. In 2014 Alanine *et al.* reported a synthetic route to substituted quinolizin-4-ones using ring-closing metathesis (RCM) as the key step in generating the ring system, followed by a palladium-catalysed oxidation to aromatise the ring system as shown in Scheme 4.1.[343] Grubbs second-generation catalyst is used to ligate the two pendant olefins in **4.15**, with X as either carbon or nitrogen. Despite detection of the partially dehydrogenated minor by-product **4.17**, good yields were reported for a range of R-group substituents.

Inspired by this strategy, a general synthetic route towards these heterocycles was designed based on the key concept of installing the 5,5-bicyclic system using an RCM reaction. RCMs are widely used in medicinal chemistry, known to reliably prepare five-membered heterocyclic rings driven by the entropically-favourable release of gaseous ethene, and the second-



Scheme 4.1: The ring-closing metathesis route to quinolizin-4-ones developed by Alanine *et al.* in 2014.

generation Grubbs catalysts are stable with high turnover numbers.[344] An outline of this proposed route is shown in Scheme 4.2.



Scheme 4.2: The proposed route to 5,5-heterocycles using ring-closing metathesis.

Unfortunately time constraints meant that exploration of these proposed syntheses within this thesis was not possible, however development of general routes towards these 5,5-bicycles remains an ongoing project within the Brennan Group.

4.4 Conclusions and Future Work

This chapter commenced by outlining two key shortcomings of the previous implementation, namely:

1. there was no way for a user to specify an important exit-vector of attachment; and

2. the source-code for the alignment and scoring algorithm is not open-source and freely distributed.

Each of these has been addressed here through the development of a unique, vector-based ligand alignment algorithm. The full implementation of this algorithm in Python using free and open-source packages, and its distribution under a permissive MIT licence on GitHub addresses the issue of 'black-box' code, and thus significantly improves the reproducibility and reliability of the codebase and the results derived from it. Furthermore, its distribution on GitHub allows contributions and suggestions from the wider scientific community, thus facilitating the continual improvement of the codebase, and widening exposure of the conceptual foundations. Details of the software and packages used in the implementation are provided in Section 8.1.

The developed package supports either one or two user-specified exit-vectors, and a method of alignment based either on iterative alignment to each probe exit-vector or a hash-based retrieval of molecules with similar exit-vector geometries was described. An alignment method based on Kabsch's algorithm minimises the RMSD between query and probe exit-vectors and ensures alignment of the ring planes, and a scoring method based on a Tanimoto similarity of the van der Waals volume overlap and a Gaussian representation of the Coulombic similarity was also described. Benchmarking of the geometry optimisation and partial charge calculations were carried out, and their validities discussed.

By way of benchmarking the algorithm, 3,5-disubstituted pyrazolopyridine (a heterocyclic motif appearing in the recently FDA-approved drugs repotrectinib and larotrectinib) was searched through HCIE.[79] From the top 10 returned results, two of them were known and

well-characterised bioisosteres of the pyrazolopyridine query, but six of these had never been used as a bioisosteric replacement previously and had the same pattern of hydrogen bond donors and acceptors as the query. It is thus proposed that these heterocycles, all of which have previously been synthesised, could be novel bioisosteres of 3,5-disubstituted pyrazolopyridine. A similar search through HCIE of 2-pyridine, the most frequently used heterocyclic motif in drugs approved by the FDA between 2013 – 2024, showed that HCIE was able to retrieve known bioisosteres from the MoBiVic library with a similar or better enrichment than the Electroshape USR algorithm whilst also providing significantly more information about the best point(s) of attachment, enabling the results to be directly passed into further computational compound enumeration and design pipelines. The top 50 returned heterocycles were inspected, and a number of 5,5-bicyclic heterocycles identified within the results. None of these 5,5-bicycles have been reported as bioisosteres of 2-pyridine, and four of them have no reported syntheses. These are thus proposed as a new class of aromatic bioisostere for this medicinally important motif.

The bioactivities of five series of small-molecule inhibitors of the NLRP3 inflammasome derived from the academic and patent literature were aligned and scored against the most potent ligand in each respective series, and the total scores correlated against the reported pIC_{50}s. The correlations were all in the range 0.43–0.75, and optimisation of the weightings of ESP and shape similarities when calculating the total score improved the correlation coefficients in all cases. This analysis demonstrated that the bioactivities of certain series are more dominated by steric shape effects, and others by electrostatic effects.

These results show that in the case of aromatic heterocyclic bioisosteric replacement, bioactivity is correlated with ESP and shape similarity, and thus it is a valid combination of metrics to use when retrieving bioisosteres from a library of aromatic heterocycles. Fur-

thermore, they demonstrate that the specific weightings of shape and ESP similarity when calculating the total can be adjusted to improve the correlation, illustrating a data-driven use of HCIE in prospective compound design.

That HCIE was able to significantly enrich the returned results with known bioisosteres of 2-pyridine compared to a random draw further illustrates the validity of this vector-based approach for the identification of bioisosteres in a virtual library. The inclusion of high-scoring heterocycles that have never previously been synthesised, alongside the successful retrieval of known bioisosteres, demonstrates HCIE's ability to not only identify established bioisosteres but also propose promising novel bioisostere candidates.

The use of HCIE in combination with the MoBiVic library is expected to expand the areas of bioisosteric chemical space available to medicinal chemists, thereby enhancing the toolkit of aromatic heterocycles for improving potency or modulating other pharmacokinetic parameters. By highlighting areas of aromatic heterocyclic chemical space that are predicted to be bioactive but remain synthetically unexplored, HCIE can help guide the development of synthetic methodologies toward these molecules, thus prioritising bioactive areas of chemical space for synthetic development.

4.4.1 Future Work

Future work in this area fits loosely into three key directions:

1. improvement of the codebase;

2. development of synthetic methodology; and

3. expansion of the alignment methodology to include saturated heterocycles.

Scientific code is not a static entity, but continually evolves to include new features and functionality, and as more widespread use uncovers bugs and unexpected behaviour. Unit and functional testing can always be improved and expanded to include new edge cases, and code can invariably be refactored to improve readability and general code quality. Future work will involve maintaining and improving the codebase to ensure that it meets the needs of current users, and is in line with current best practice in scientific software engineering.

As aforementioned in Section 4.3.3.1, a general synthetic route towards these 5,5-bicyclic heterocycles has been proposed but time restrictions meant it has remained unexplored. A significant area of future work is to explore the feasibility of this proposed route, including optimising the reaction conditions and testing its scope. When this is optimised, analogues of literature inhibitors (for which reliable assays exist and are straightforward to set up and run) where 2-pyridine moieties are substituted for these 5,5-bicycles will be synthesised, and their bioactivities determined.

Finally, although the principal focus of this thesis is the exploration and discovery of aromatic heterocyclic biosisosteres, the region of saturated or partially saturated heterocyclic chemical space is larger, and certainly holds significant bioisosteric potential. There has been a recent surge of interest in the discovery of sp^3-rich bioisosteres of benzene, but there is not as yet a means of systematically searching and proposing the structures of these aliphatic ring systems.[201] An important line of future work is the expansion of the methodology described here, and the MoBiVic library, to include saturated and partially saturated heterocycles. The inclusion of these sp^3 centres into the methodology will significantly expand the diversity

of proposed bioisosteres, and could inspire novel therapeutic scaffolds that push beyond the constraints of traditional aromatic systems.

5 Efforts Towards the Synthesis of Novel VEHICLe Heterocycles

The results described in this chapter contributed to the following publication, and are therefore based in part on work published therein.

Transfer Learning for Heterocycle Retrosynthesis, E. Wieczorek, J. W. Sin, S. Tanovic, M. T. O. Holland, L. Wilbraham, V. Sebastián-Pérez, A. Bradley, D. Miketa, P. E. Brennan, F. Duarte *J. Chem. Inf. Model.*. DOI: 10.1021/acs.jcim.4c02041 **2025**.

5.1 Introduction

As highlighted in Chapter 1, at the time of publication in 2009 Pitt *et al.* estimated that around 1300 of the VEHICLe heterocycles were within the reach of modern synthetic methodologies but were as yet unsynthesised.[256] This chapter describes efforts made to synthesise several of these heterocycles, following disconnections suggested by a ring-breaking transformer machine learning (ML) model specifically trained on a dataset of heterocycles within the Brennan and Duarte Groups.[345]

Despite the widespread prevalence of heterocycles in bioactive molecules (Marshall *et al.* found that 82% of drugs approved by the FDA from 2013 - 2023 contained at least one nitrogen heterocycle), the proportion of the known chemical reactions that lead to their generation is small, with a 2020 study finding that only 4.5% of a dataset of reactions derived from the US Patent Office data and 5.8% of the Reaxys dataset were ring-forming reactions.[79,346] Computer-aided synthesis planning (CASP) tools exist, but they are often trained on general reaction datasets of which ring forming reactions typically make up a very small fraction, and the principal focus of recent synthetic efforts has been on derivatising rings rather than forming them.[347,348] Work in the Duarte Group[a] sought to address this by using domain adaptation to improve the retrosynthetic performance of a sequence-to-sequence (*seq2seq*) molecular transformer ML architecture on heterocycle formation tasks.[345]

A molecular transformer is an ML model based on the transformer architecture first proposed by Vaswani *et al.* in 2017, and trained on large datasets of chemical reactions to predict retrosynthetic disconnections.[349,350] As opposed to following template-based rules for retrosynthetic disconnection, *seq2seq* transformers treat molecules (in the form of SMILES strings) as tokens and thus treat retrosynthesis in a similar manner to language translation using self-attention, a technique which allows the model to focus on the most relevant parts of the input molecule when predicting disconnections. By leveraging self-attention, the molecular transformer can identify key functional groups and reaction centres without relying on predefined heuristics, making it more adaptable to predicting novel transformations.

Taking a molecular transformer model pre-trained on a large corpus of reaction data (over 1.1 million reactions, the *General* dataset), the Duarte Group assembled a dataset of 165

[a]The work on dataset formation and ML model training described here was carried out by Ewa Wieczorek and Joshua Sin.

216 ring-forming reactions from academic literature and patent data (the *Ring* dataset), and used a mixed fine-tuning approach to update the model to improve its performance on ring forming disconnections (see Figure 5.1).[350]

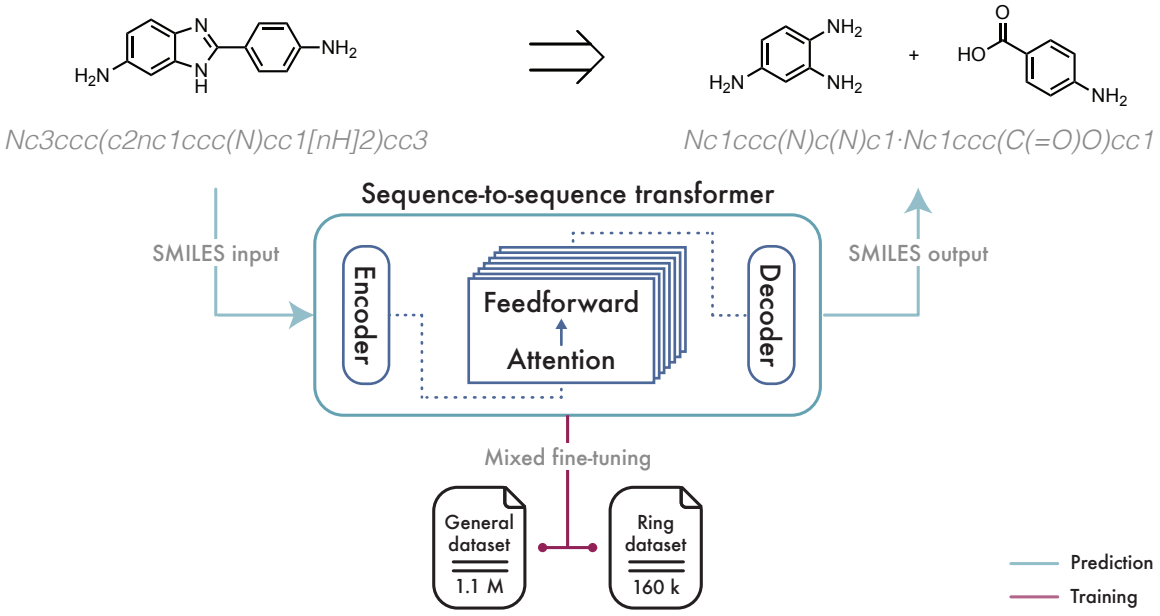


Figure 5.1. An illustration of the *seq2seq* transformer architecture and the mixed fine-tuning approach taken by Wieczorek *et al.*. This figure is adapted from those in Reference [345].

Mixed fine-tuning is an ML training approach that involves updating a pre-trained ML model using a combined dataset of domain-specific data (in this case ring-forming reactions from the *Ring* dataset) and a subset of the original dataset used for initial training. This adapts the model to perform better on reactions from the ring-forming domain, whilst preventing catastrophic forgetting of the original training data, which encodes important general chemistry knowledge.

To investigate whether this fine-tuned model could be used to design synthetic routes towards previously unsynthesised aromatic heterocycles, synthetic routes were predicted for each of the VEHICLe heterocycles reported as being unsynthesised in the original publication. These were then filtered to exclude disconnections where the model predicted a low probability of

success (< 0.8), and six heterocycles were manually selected from those remaining based on ease-of-accessibility of the starting materials and reagents, and the feasibility of the proposed disconnection. These are illustrated in Figure 5.2.

Both **a** and **b** use triphosgene as a $O{=}C^{2+}$ synthon to form a 5,6 bicyclic ring from a pendant carboxylic acid and acyl thiol or amine respectively. Interestingly the model included THF as a reactant in both cases, presumably as a solvent, but only in **a** was a base (triethylamine) included in the proposed reaction. **c** and **d** both make use of the condensation of a heteroaryl aldehyde with ethyl carbazate to form a hydrazone intermediate, which then further cyclises with the loss of EtOH to form the bicyclic products. Two further reactions of different classes were also selected, with **e** using hydroxylamine to cyclise furan-3-methoxy-4-carbaldehyde **5.9** via an oxime intermediate, and **f** using 1,1,3,3-tetraethoxypropane (1,1,3,3-TEP) as a masked dialdehyde to cyclise **cytosine**. All have high predicted confidences, with all but **e** having a probability > 0.9.

Although the model provides the retrosynthetic disconnections, and occasionally suggests solvent molecules as part of these (*vide supra*), in general it gives no information on suggested experimental conditions for the transformations (e.g. time, temperature, or solvent if not included in the disconnection). Furthermore, the model does not provide links or citations to the original literature references from which the disconnections were derived or that serve as precedent for the suggested transformations. Therefore, unless otherwise stated, initial starting conditions for these cyclisations were derived from chemical intuition, guided where available by literature precedent for similar reactions.

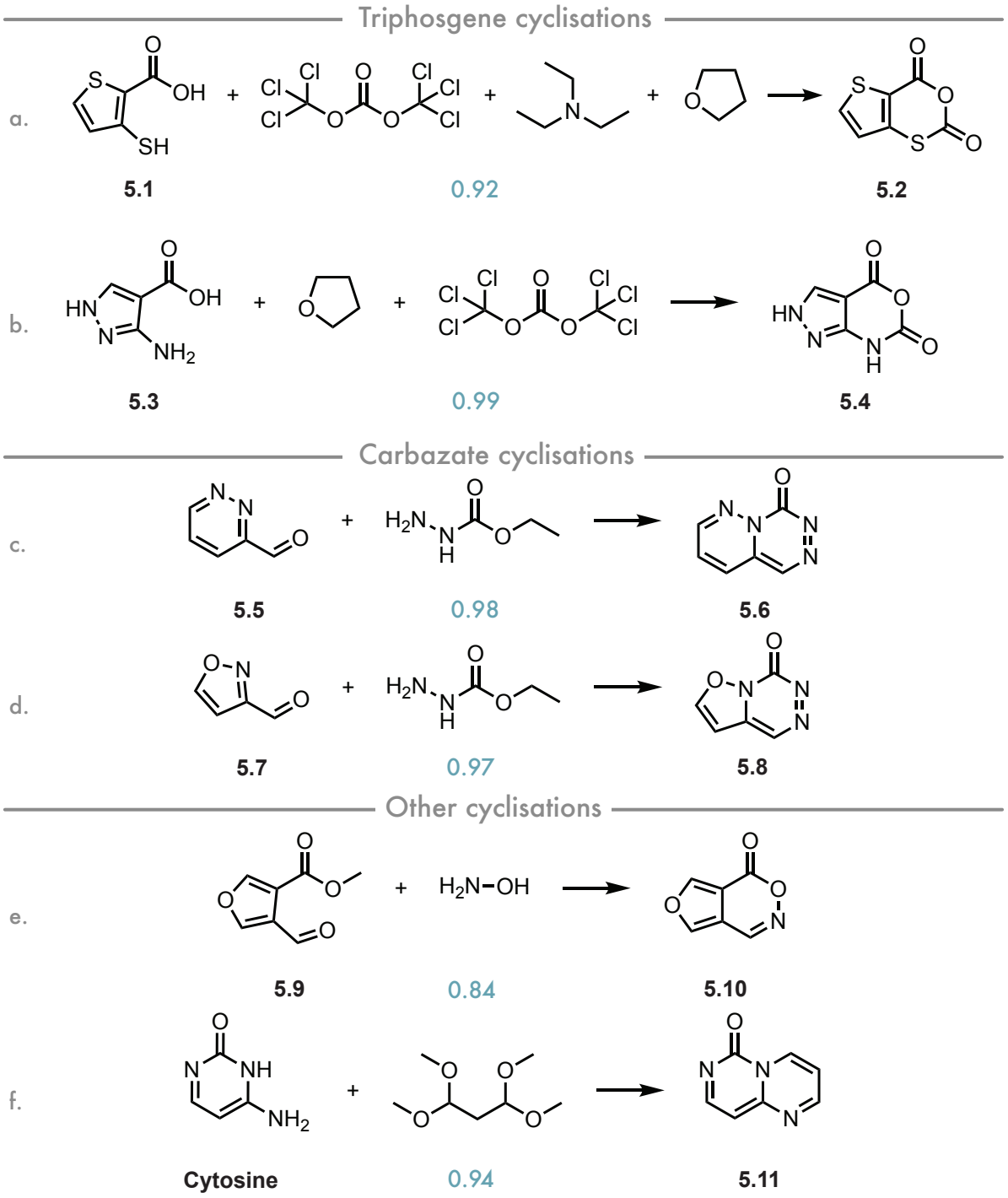


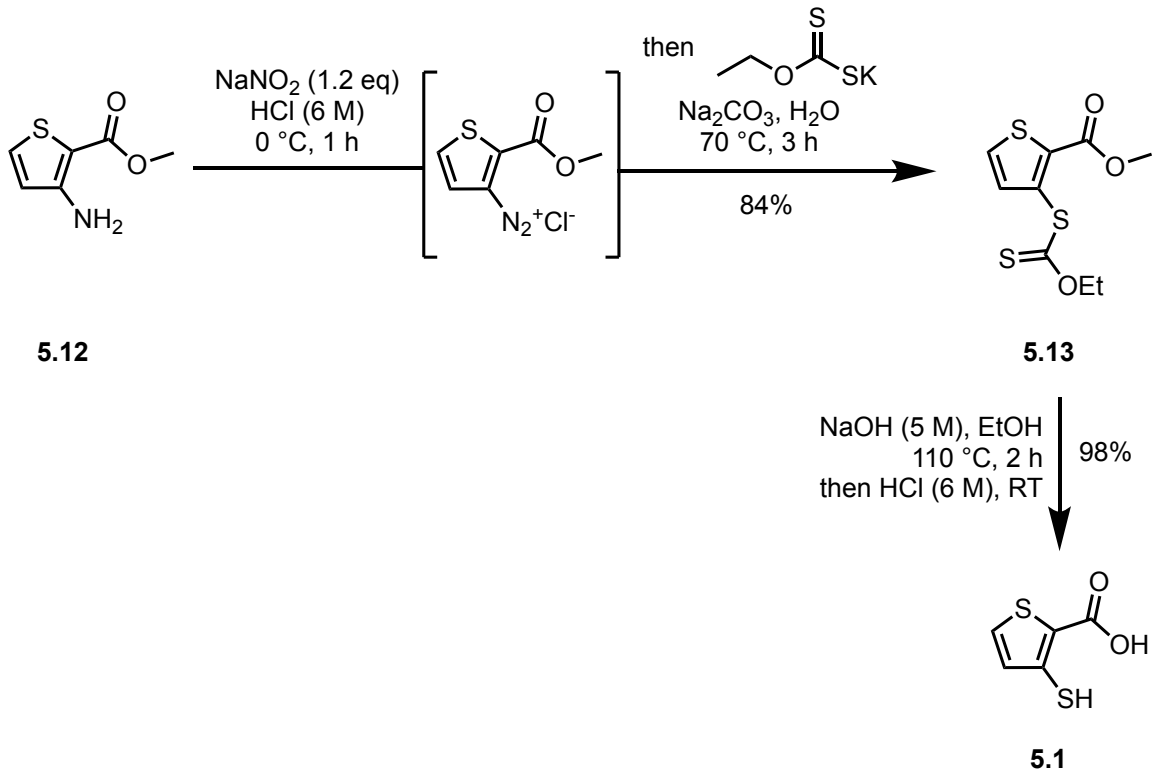
Figure 5.2. The VEHICLe heterocycles and their machine-learning disconnections chosen for synthetic evaluation. The number in blue represents the probabilistic confidence of the model in the given disconnection.

5.2 Triphosgene Cyclisations

5.2.1 Towards the Synthesis of 5.2

Initially the starting material **5.1** was prepared from commercially available amino ester **5.12** following the literature precedent of Hu *et al.*, as illustrated in Scheme 5.1.[351] Nitrous acid was generated *in situ* from sodium nitrite and hydrochloric acid, and this used to diazotise **5.12**, giving the diazo intermediate which was not isolated but immediately treated with potassium ethyl xanthogenate and sodium carbonate in a nucleophilic aromatic substitution reaction (the excellence of the diazo leaving group here making up for the poor nucleophilicity of the xanthogenate anion), to gave xanthogenate **5.13** in 84% yield. This was telescoped without further purification, and the methyl ester and xanthogenate hydrolysed with ethanolic sodium hydroxide, the resulting malodorous solution acidified, and **5.1** isolated in overall 82% yield.

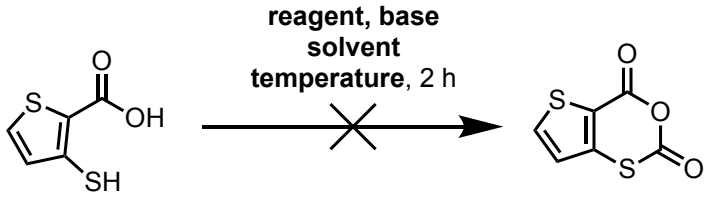
Taking the view that the disconnections proposed by the model were more important than the precise conditions proposed, and wishing to avoid the use of highly toxic triphosgene, carbonyldiimidazole (CDI) was initially screened as a safer source of the $O{=}C^{2+}$ synthon (see Table 5.1). Inspired by similar transformations reported by Sahner *et al.* and Zhang *et al.*, the reaction was initially attempted in THF at room temperature (entry 1) but yielded just starting material, as did repeats at 65 °C both with and without base (entries 2 and 3).[352,353] Attempts in dioxane and DMA (entries 4-6) both gave no conversion, even at elevated temperature (entry 7). The cyclisation was then attempted in triphosgene as per the conditions proposed by the model. A Reaxys search for triphosgene cyclisations revealed a significant literature precedent for DCM solvent, thus reaction at room temperature in DCM with triethylamine base was attempted (entry 8) immediately yielding a red precipitate,



Scheme 5.1: The synthesis of 3-mercaptothiophene-2-carboxylic acid **5.1**.

whose quantity increased over the course of the two hour reaction. This was isolated by filtration, but was insoluble in all solvents except DMSO. LCMS and TLC analysis showed no discernible mass, and ^{1}H NMR spectroscopy of a concentrated sample in d^6-DMSO showed no significant peaks across the typical proton chemical shift range. Alternative ionisation techniques (electrospray ionisation, atmospheric pressure chemical ionisation, and electron ionisation) failed to record any relevant ions. It is proposed that this red precipitate is a result of the triphosgene-mediated polymerisation of **5.1**, thus further attempts at reduced **5.1** concentration with the intention of favouring cyclisation over polymerisation failed to show any conversion to the desired product. This red precipitate was also observed in the reaction with triphosgene in THF at 65 °C (entry 11). Due to time constraints, attempts at this cyclisation were not continued.

Table 5.1. Results of the screening of conditions for the attempted cyclisation of mercaptothiophene **5.1**. *Reagents and conditions:* **5.1** (1.0 mmol), reagent, base, solvent, temperature, 2 h.



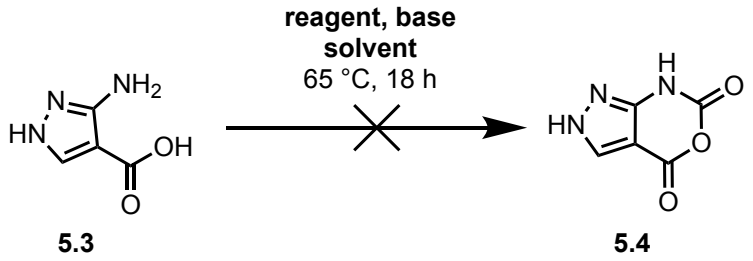
5.1 **5.2**

Entry	Reagent (equiv.)	Base	Solvent	Temperature
1	CDI (1.5)	Et_3N	THF	R.T.
2	CDI (1.5)	Et_3N	THF	65 °C
3	CDI (1.5)	—	THF	65 °C
4	CDI (1.5)	Et_3N	dioxane	R.T.
5	CDI (1.5)	Et_3N	dioxane	65 °C
6	CDI (1.5)	Et_3N	DMA	65 °C
7	CDI (1.5)	Et_3N	DMA	120 °C
8	triphosgene (1.0)	Et_3N	DCM	R.T.
10	triphosgene (1.0)	—	THF	R.T.
11	triphosgene (1.0)	Et_3N	THF	65 °C

5.2.2 Screening of Conditions for the Synthesis of 5.4

5.3 was sourced commercially, and conditions for its cyclisation screened, as outlined in Table 5.2. Following the unsuccessful attempts at accessing **5.2**, it was decided to screen conditions. An initial solvent screen across a range of solvent polarities (entries 1-5) was therefore conducted. In each case a canary-yellow precipitate formed in the reaction vessel, but attempts to dissolve this in any solvent except DMSO failed, and NMR spectroscopy and LCMS studies were inconclusive. It was therefore assumed that the polymerisation of the starting material, akin to that outlined in Section 5.2.1, was occurring in this instance too. Attempts without base (entry 6), or with triphosgene (entry 7) also only gave canary-yellow precipitate.

Table 5.2. The conditions screened for the cyclisation of aminopyrazole carboxylic acid **5.3**. *Reagents and conditions:* **5.3** (0.8 mmol), reagent, base (5 eq.), solvent (8.5 mL), 65 °C, 18 h.



5.3 **5.4**

Entry	Reagent (equiv.)	Base	Solvent
1	CDI (1.5)	Et_3N	THF
2	CDI (1.5)	Et_3N	dioxane
3	CDI (1.5)	Et_3N	DMF
4	CDI (1.5)	Et_3N	toluene
5	CDI (1.5)	Et_3N	MeCN
6	CDI (1.5)	—	THF
7	triphosgene (1.0)	—	THF

Taken with results for **5.2** outlined above, it was decided that the starting materials are too prone to polymerisation, and further attempts at these disconnections were not pursued.

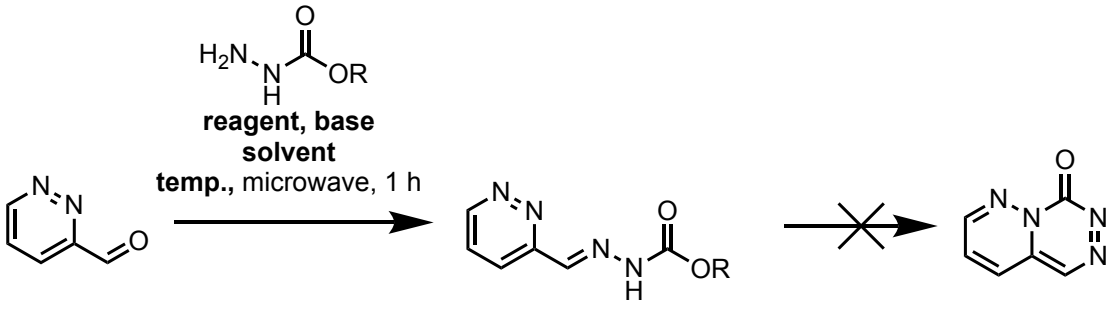
5.3 Carbazate cyclisations

To test the reactions proposed in Figure 5.2**c** and **d**, heteroaryl aldehyde starting materials **5.5** and **5.7** were sourced commercially. Given the limited literature precedent, it was decided initially to screen a variety of conditions for the cyclisation of **5.5** with ethyl carbazate (see Table 5.3), including a solvent screen across a range of solvent polarities (entries 1-5, 8-9) and elevated temperatures (entries 6-7, 10), with a view to identifying a starting point for further optimisation. All of these reactions showed formation of hydrazone intermediate **5.14**, but no evidence by TLC, LCMS, or NMR spectroscopy of the cyclised product. The conversion of **5.5** to **5.14** was not quantified in these test reactions, but observed qualitatively by LCMS

in all cases. To test whether the lack of cyclisation of the hydrazone was related to the poor nucleophilicity of the pyridazine nitrogen, the effect on the outcome of the addition of catalytic quantities of oxophilic Lewis acid scandium(III) triflate was investigated (entries 11 and 12).[354] It was theorised that the coordination of the carbazate carbonyl oxygen to the Sc(III) centre would enhance the electrophilicity of the carbonyl carbon, thus favouring the nucleophilic attack of the pyridazine nitrogen and promoting cyclisation. Unfortunately no cyclised product was observed, either at room temperature or at 100 °C.

Further conditions were then screened, this time using methyl carbazate as a more reactive surrogate for ethyl carbazate (the reduced steric hindrance and lower electron-donating inductive effect afforded by the methyl group is expected to render the carbonyl carbon more electrophilic; entries 13-15). Initial high temperature heating in DMA and DMSO (entries 13-14) showed methoxyhydrazone formation, but no evidence of cyclised product. To investigate whether thermal barriers were preventing cyclisation, heating at 210 °C in DMSO was attempted (entry 14) and showed black, tarry residue on the microwave vial, suggesting degradation of starting materials or intermediate products. To investigate whether the addition of base would abstract the hydrazone proton and drive cyclisation, non-nucleophilic bases NaO^tBu, NaH, NaHMDS, and also NaOH were screened. Initial test reactions (entries 15-16) formed a brown precipitate, which was identified as the sodium salt of the deprotonated hydrazone **5.14**. To prevent salt formation, sodium-chelating crown ether 15-crown-5 was added (entries 17-18, 21-22), but still no cyclised product was observed. Further attempts to promote cyclisation with oxophilic Lewis acids scandium(III) triflate and ytterbium(III) triflate (entries 19-22) also did not yield cyclised product. Final attempts to avoid hydrazone salt precipitation involved exploiting the low solubility of sodium bromide in acetonitrile through a salt metathesis reaction with the bulky, non-coordinating tetrabutylammonium

Table 5.3. Results of the screening of conditions for the attempted cyclisation pyridazine-3-carbaldehyde **5.5** with alkoxy carbazate. *Reagents and Conditions:* **5.5** (0.1 mmol), Lewis acids (0.2 eq. if used), base (2.1 eq. if used), crown ether (2.1 eq. if used), tetrabutylammonium bromide (TBAB, 2.1 eq. if used). [a]This reaction showed evidence of material degradation.



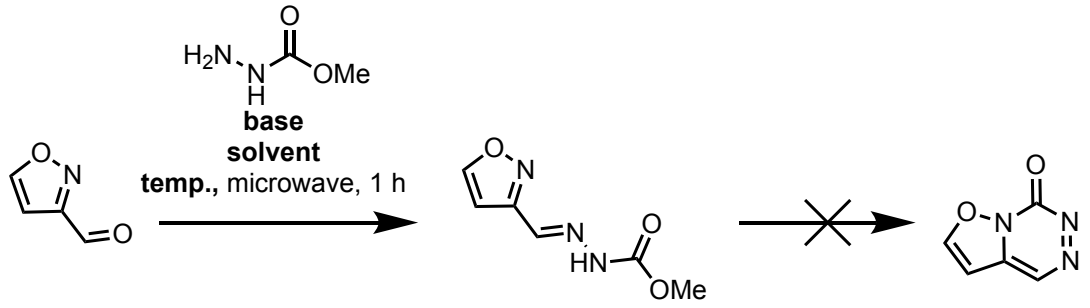
5.5 **5.14** **5.6**

Entry	R	Reagent	Base	Solvent	Temp.
1	Et	—	—	toluene	RT
2	Et	—	—	dioxane	RT
3	Et	—	—	$CHCl_3$	RT
4	Et	—	—	EtOAc	RT
5	Et	—	—	AcOH	RT
6	Et	—	—	AcOH	100 °C
7	Et	—	—	AcOH	200 °C
8	Et	—	—	EtOH	RT
9	Et	—	—	MeCN	RT
10	Et	—	—	MeCN	100 °C
11	Et	$Sc(OTf)_3$	—	MeCN	RT
12	Et	$Sc(OTf)_3$	—	MeCN	100 °C
13	Me	—	—	DMA	120 °C
14[a]	Me	—	—	DMSO	210 °C
15	Me	—	NaOtBu	MeCN	80 °C
16	Me	—	NaOH	MeCN	80 °C
17	Me	15-crown-5	NaOtBu	MeCN	80 °C
18	Me	15-crown-5	NaH	MeCN	80 °C
19	Me	$Sc(OTf)_3$	NaH	MeCN	80 °C
20	Me	$Yb(OTf)_3$	NaH	MeCN	80 °C
21	Me	$Sc(OTf)_3$, 15-crown-5	NaH	MeCN	80 °C
22	Me	$Yb(OTf)_3$, 15-crown-5	NaH	MeCN	80 °C
23	Me	TBAB	NaH	MeCN	80 °C
24	Me	TBAB	NaOtBu	MeCN	80 °C
25	Me	TBAB	NaHMDS	MeCN	80 °C

cation (entries 23-25). Unfortunately cyclised product was still not observed. It was decided at this point to abandon further attempts at this fused pyridazine disconnection.

As the isoxazole-3-carbaldehyde starting material **5.7** was commercially available, it was decided to make initial attempts at screening conditions for the cyclisation reaction proposed in Figure 5.2**d**. The results of this screening are displayed in Table 5.4.

Table 5.4. The screened conditions for the cyclisation of isoxazole-3-carbaldehyde **5.7**. *Reagents and conditions:* **5.7** (0.2 mmol), methyl carbazate (1.6 eq.), base (1.1 eq.), heated in a microwave at specified temperature for 1 h.



5.7 **5.15** **5.8**

Entry	Base	Solvent	Temperature
1	—	DMA	RT
2	—	DMA	90 °C
3	—	DMA	120 °C
4	—	MeCN	120 °C
5	NaO^tBu	MeCN	150 °C

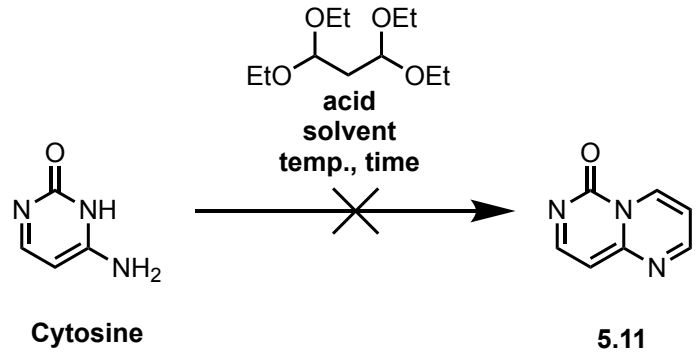
Reactions in DMA at a range of temperatures (entries 1-3) again gave qualitative evidence (by LCMS analysis) of formation of hydrazone **5.15**, but no evidence of cyclised product was observed. Attempts in acetonitrile with or without non-nucleophilic base (entries 4-5) also gave no observable cyclisation product. At this stage, it was decided to make no further attempts at the carbazate-mediated cyclisations.

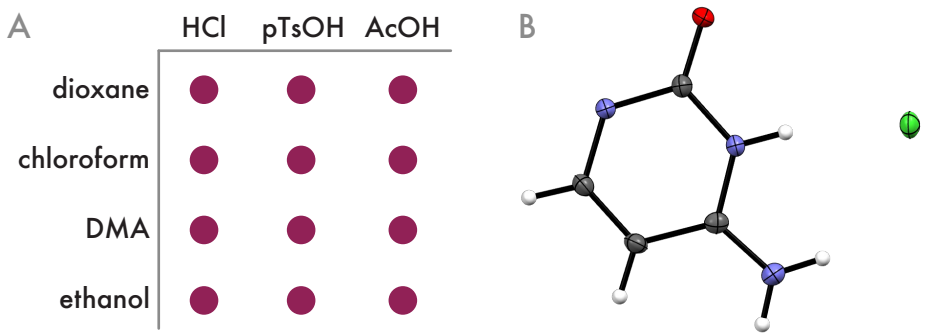
5.4 Cytosine-Derived Cyclisations

The starting material for the disconnection in Figure 5.2**f** was the nucleotide base cytosine, which is readily commercially available. The proposed disconnection involved a sequence of two imine-like condensations from the amino and urea-like nitrogens of cytosine onto the masked dialdehyde 1,1,3,3-TEP. Initially a solvent and acid screening panel was attempted at 85 °C over 16 hours to identify a starting set of conditions for further optimisation, and the results of this shown (with a red dot indicating no evidence of cyclised product) in Panel A of Table 5.5. None of the screened acid-solvent combinations gave any evidence of the cyclised product by TLC, LCMS, or ^{1}H NMR spectroscopy. Further screenings in ethanol across a variety of temperatures (entries 1-6), with microwave heating (entry 6), and with a large excess of 1,1,3,3-TEP (entry 5) did not give any evidence of cyclisation. However, upon cooling, large rhombic crystals were observed in the reaction vessel for entry 6. X-ray diffraction of these crystals revealed the HCl salt of the cytosine starting material, for which the derived structure is shown in Panel B of Table 5.5. Screening of weaker tosic acid in a variety of solvents under microwave irradiation (entries 7-13) also failed to give any observable cyclised product. Finally a procedure inspired by similar reactivity documented by Sun *et al.* involving acetic anhydride and acetic acid solvent at elevated temperature (entry 14) was attempted, but again this did not yield any cyclised product.[355]

To investigate the effects of electronically or sterically altering the cytosine starting material, and to make the starting material (and intended product) more lipophilic with the hope of making it easier to detect and isolate, 5-bromocytosine **5.16** was prepared from cytosine by electrophilic aromatic substitution with *N*-bromosuccinimide in DMF at 68% yield (Scheme 5.2). A Suzuki coupling with phenylboronic acid in 1,4-dioxane with XPhos Pd second generation catalyst and aqueous tribasic potassium phosphate gave 5-phenylcytosine

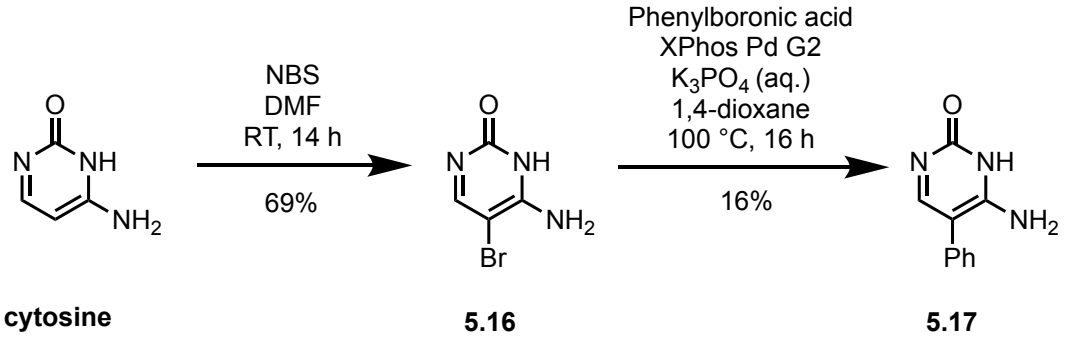
Table 5.5. Results of the screening of conditions for the cyclisation of **cytosine** with 1,1,3,3-tetraethoxypropane (1,1,3,3-TEP). Panel A shows the results of a solvent and acid panel screening of conditions *reagents and conditions:* **cytosine** (0.2 mmol), 1,1,3,3-TEP (2 eq.), solvent (1.5 mL), acid (10 µL/15 mg), 85 °C, 16 h. Panel B shows the crystal structure obtained from entry 4 after cooling. [a] Reaction heated in microwave for 30 minutes with 15 eq. of 1,1,3,3-TEP. [b] 23 eq. of 1,1,3,3-TEP. [c] 2.5 eq. 1,1,3,3-TEP, 10 eq. of Ac_2O, AcOH (1.8 mL).





Entry	Acid	Solvent	Temperature (°C)	Time
1	—	EtOH	90	2 h
2	HCl	EtOH	40	2 h
3	HCl	EtOH	60	2 h
4	HCl	EtOH	90	2 h
5[a]	HCl	EtOH	100	45 mins
6[a,b]	HCl	EtOH	100	15 mins
7[a,b]	pTsOH	EtOH	100	30 mins
8[a]	pTsOH	EtOH:MeCN (3:1)	100	1 h
9[a]	pTsOH	MeCN	100	15 mins
10[a]	pTsOH	MeOH	100	15 mins
11[a]	pTsOH	H_2O:EtOH (3:1)	100	15 mins
12[a]	pTsOH	THF	100	15 mins
13[a]	pTsOH	toluene	130	45 mins
14[a,c]	Ac_2O	AcOH	110	2 h

5.17 in 16% yield. There was significant evidence of debrominated cytosine in the reaction mixture, however **5.17** was isolated in sufficient quantity for further investigations, and as such in the interests of time the coupling reaction was not further optimised.



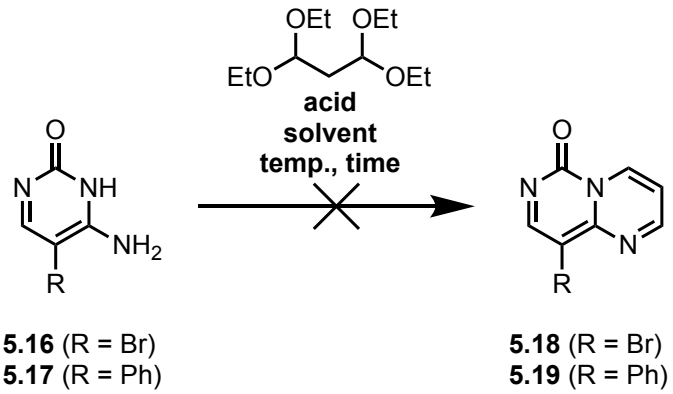
Scheme 5.2: The synthesis of bromocytosine **5.16** and phenylcytosine **5.17** from cytosine.

Table 5.6 shows the results of screening both bromocytosine **5.16** and phenylcytosine **5.17** for cyclisation with 1,1,3,3-TEP. As no cyclisation had been observed under any of the attempted conditions, or for any of the modified cytosine substrates, further attempts at validating this disconnection were not pursued.

5.5 Furoxazinone Synthesis

The starting materials for the forward reaction proposed in Figure 5.2**e** were not commercially available, and so a synthetic route was designed from commercially available dimethyl-3,4-furandicarboxylate **5.20** was designed, and is outlined in Scheme 5.3. An ester hydrolysis of the diester starting material with 2 M potassium hydroxide solution in methanol at room temperature for one and a half hours gave monohydrolysed product **5.21** in 75% yield.[356] Reduction of the carboxylic acid with dimethylsulfide adducted borane in THF gave the alcohol product **5.22** in quantitative yield, and a Swern oxidation afforded the aldehyde **5.9** in 35% yield, thus giving an overall yield of 26% across the three steps.

Table 5.6. Screened conditions for the cyclisation of **5.16** and **5.17** with 1,1,3,3-TEP. *Reagents and conditions:* **5.16** (1.3 mmol), 1,1,3,3-TEP (1.1 eq.), AcOH (10 mL) *or* **5.17** (0.1 mmol), acid (0.3 eq.), solvent (1 mL).[a] Heated in a microwave.

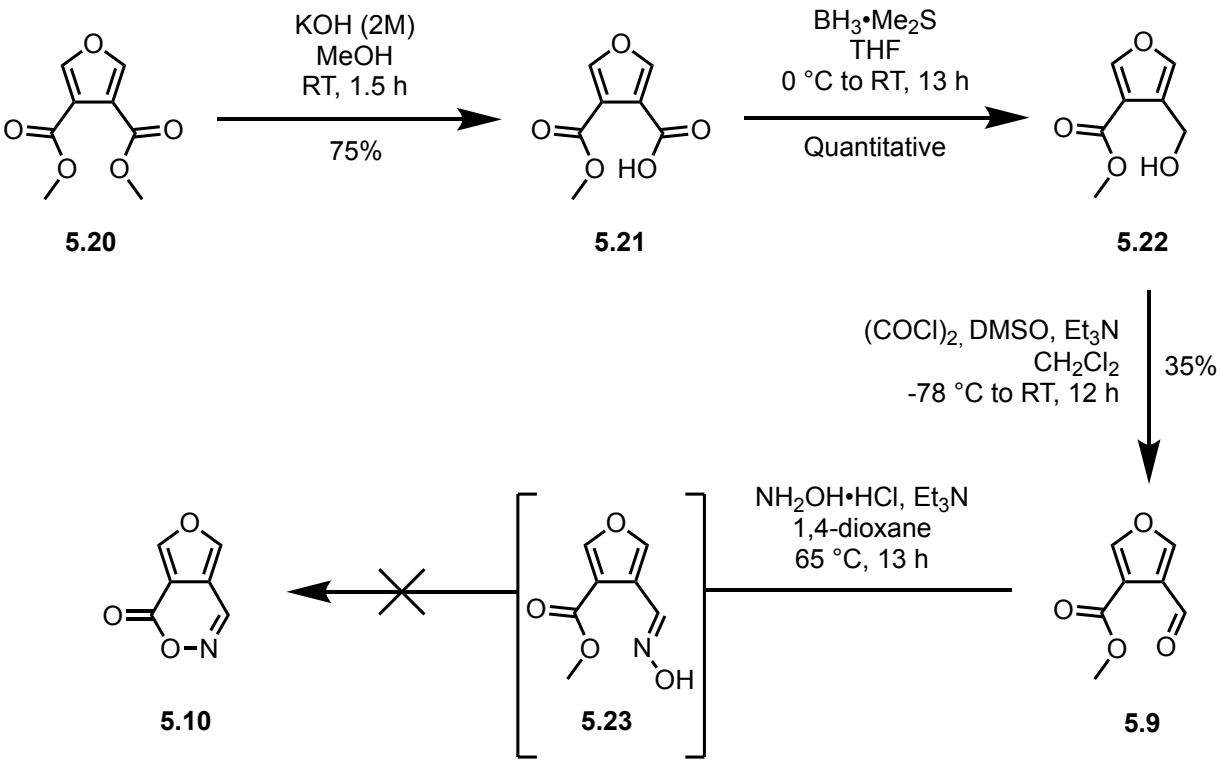


5.16 (R = Br)
5.17 (R = Ph)

5.18 (R = Br)
5.19 (R = Ph)

Entry	R	Acid	Solvent	Temperature (°C)	Time (h)
1	Br	AcOH	AcOH	110	16
2[a]	Ph	pTsOH	EtOH	130	1.5
3[a]	Ph	pTsOH	DMA	130	1.5

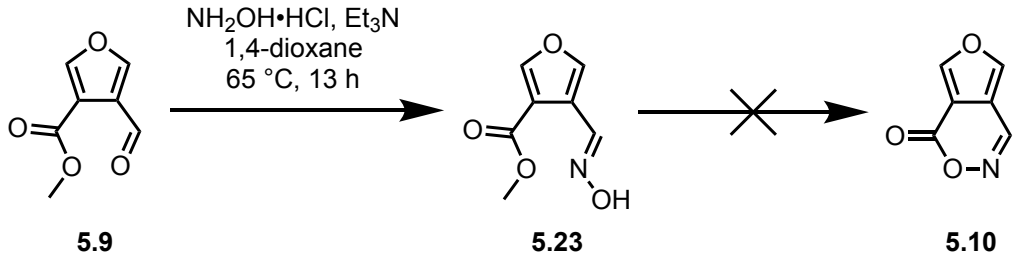
An initial attempt at testing the ML-proposed disconnection is given in Scheme 5.4. Hydroxylamine hydrochloride, triethylamine, and **5.9** were taken up in dioxane (it was known that both starting materials were soluble in dioxane, thus this was chosen as an initial solvent) and stirred at 65 °C for 13 hours. LCMS and TLC analysis showed formation of oxime derivative **5.23**, which was not quantified, but no cyclised product was observed in the reaction mixture by LCMS or ^{1}H NMR spectroscopy. Although further screening is required to ascertain the full feasibility of the proposed disconnection, time constraints meant that further attempts at this cyclisation were not pursued.



Scheme 5.3: The synthesis of aldehyde **5.9** from commercially available **5.21**.

5.6 Conclusions and Future Work

This chapter aimed to develop cyclisation reactions for VEHICLe heterocycles that had never previously been synthesised by testing the disconnections proposed by a molecular transformer ML model where mixed fine-tuning had been used as a domain adaptation strategy to improve its performance on heterocyclic ring formation reactions. Six disconnec-



Scheme 5.4: The attempted cyclisation of **5.9** with hydroxylamine. Oxime **5.23** was identified, but no further cyclisation to **5.10** was observed.

tions on VEHICLe heterocycles proposed by the model were selected, and attempts towards validating them were described.

The triphosgene cyclisations **a** and **b** produced sparingly soluble, coloured precipitates in both cases, which is likely to be polymerised starting material. This suggests that the stability of the cyclised product is too low for these disconnections to be fruitful, with polymerisation preferred over cyclisation despite the entropic penalty this entails. The results of these cyclisation attempts informed in part the rules described in Chapter 3 for filtering the MoBiVic library, specifically with regards to the exclusion of thioester and anhydride-like functionalities.

Each of the carbazate cyclisations **c** and **d** showed qualitative conversion to the hydrazone intermediates, but cyclisation to the desired products proved elusive in both cases. Attempts at promoting cyclisation by enhancing the electrophilicity of the carbonyl carbon with oxophilic Lewis acid, or by deprotonating the hydrazone proton and sequestering the metal cation to prevent salt formation failed to yield any desired product. The geometry of the hydrazone is an area for concentration of future research. Hydrazones can exist in both *E*- and *Z*- isomers, with the *E*-isomer likely to be lowest in energy due to reduced steric interactions.[357–359] The trans configuration of the intended nucleophilic atom and the electrophilic hydrazone carbonyl carbon in this isomer renders the two reacting centres far apart, which could potentially explain the failure of the reaction to progress beyond the hydrazone intermediate. Hydrazones have found use in the literature in recent years as molecular photoswitches, with certain visible wavelengths used to photoisomerise the hydrazone double bond and thus change the molecule's geometry.[360–362] A potential route to **5.6** and **5.8** could involve formation of the hydrazone, followed by some form of photoisomerisation to access the *Z*-isomer, which would hopefully bring the nucleophilic and electrophilic centres in close

enough proximity to react. Calculations to determine the energy, and thus the wavelength required to excite this transition, are currently ongoing. It is also likely that the geometry of the hydrazone can be influenced by substitution at the hydrazone carbon, with more sterically demanding substituents likely to favour formation of the desired *cis* isomer.

The attempts at cyclising **cytosine** were hampered by the polarity of the starting materials and the predicted polarity of the product, which made LCMS identification and isolation of any reaction (side-)products challenging. Attempts to mitigate this by derivatising the **cytosine** starting material did not lead to any successful cyclisations, but made monitoring and diagnosing the reactions more straightforward. It is likely in all of the disconnections investigated in this Chapter that the low molecular weight and high polarity of the desired products will have made reaction monitoring, and possibly the success of the reaction, difficult to achieve. Future attempts at reactions such as these, and at the synthesis of further VEHICLe heterocycles, should focus on more highly derivatised and functionalised scaffolds, such that the lipophilicities and molecular weights of the starting materials and products are in a range that makes LCMS and TLC tracking of the reactions straightforward. Furthermore, functionalised aromatic rings (for example halides) are then able to undergo further reaction, thus increasing the relevance of the novel created heterocycles.

Finally, the cyclisation of **5.9** to give **5.10** was under-investigated due to time constraints. Further conditions, including solvents and temperatures, should be screened to determine the feasibility of the proposed reaction. Like the hydrazones outlined above, oximes such as the observed intermediate **5.23** also form geometric isomers, and it is likely that the energetically favoured *E*-oxime is not the isomer that is able to attack into the ester carbonyl and cyclise.[363] Although the conditions for isomerisation of oximes is sparsely reported in the literature, there is some evidence to suggest that *E*-*Z* isomerisation is thermally accessible.[363,364] As

such, future work should focus on exploring the thermal conditions required to isomerise the oxime and investigate whether this is sufficient to promote cyclisation, ideally on a more functionalised starting scaffold.

Although these studies ultimately failed to achieve the primary goals of developing syntheses of novel VEHICLe heterocycles or validating the disconnections proposed by the ML model, they highlighted a few key considerations that should be evaluated before undertaking exploratory synthetic studies of this kind. Key amongst these is the consideration of the most favourable geometric isomers of any intermediates, and how these might be controlled to promote desired reactivity, and the size and polarity of the starting materials and products; reactions with more highly functionalised starting materials are easier to monitor and diagnose.

6 Translating HCIE to Practice: Experimental Applications in Medicinally Relevant Case Studies

The proof of the pudding is in the eating.

English proverb.

Having illustrated previously the development and refinement of HCIE, this chapter outlines its practical application in ongoing medicinal chemistry projects within the Centre for Medicines Discovery (CMD). Both the initial version of HCIE (as described in Chapter 2) and the second generation (Chapter 4) were made available to scientists within the CMD to use as part of their ongoing projects to test its applicability, and to generate data & user-feedback to further refine and improve the implementation. Described herein are two ongoing projects where computational techniques including HCIE made contributions towards the results and progress of the project.

Section 6.1 outlines efforts within the CMD to develop selective chemical probes for NUDT14, and describes HCIE's role in the design and the synthesis of a compound to probe the effects of the electronics of a heterocyclic ring on the efficiency of $\pi - \pi$ stacking interactions in the binding pocket. Whilst the primary focus of this project was experimental, HCIE provided computational insights that complemented and informed aspects of this work, and will guide future efforts. Section 6.2 describes previous virtual screening efforts to identify potent and metabolically-stable inhibitors of the histone methyltransferase (HMT) enzymes SUV4-20 H1 and H2 as part of an ongoing CMD project on the rare disease Friedreich's Ataxia. Compounds were designed based on the results of the virtual screening efforts, and their synthesis and initial assay results are described. The role of HCIE in ongoing and future work modulating the 6,7-dichlorophthalazine core is outlined.

6.1 Chemical Probes for NUDT14

The CMD and the Brennan Group have a long-standing interest in the development of chemical probes for medicinally important targets.[365–372] Chemical probes are small-molecule modulators of a target protein's function that can be used to investigate its function, or its tractability as a pharmaceutical target (for example its ability to modulate a phenotype or the safety of inhibiting it).[138,373] In order to classify as a high-quality probe, and thus be more likely to generate reliable and reproducible experimental results, a compound must meet the following criteria:[138,374]

- A minimum *in vitro* potency of 100 nM.
- Greater than 30× selectivity over sequence-related proteins from the same target family.

- Be profiled against an industry standard selection of pharmacologically relevant 'off-target' proteins.

- Have demonstrated on-target cellular effects below 1 µM.

Recently, effort in the CMD has been directed towards discovering chemical probes for proteins in the NUDIX family.[375] These enzymes catalyse the hydrolysis of nucleoside diphosphates bonded to other moieties (the X in NUDIX), however despite the identification of over 20 mammalian members of this family, the distinct biological and cellular roles of these proteins are largely unexplored.[375–377] NUDT5 has been suggested as a key enzyme in nuclear ATP synthesis, and is implicated in tumour cell proliferation, with the inhibition of NUDT5 activity shown to impair the growth of breast cancer cells in various models.[378–380] It was shown in 2017 that both NUDT14 and NUDT5 are able to hydrolyse the important cell-signalling molecules ADP-ribose and ADP-glucose, however the precise cellular role of NUDT14 is not well-understood.[377]

Seeking to shed light on the precise biological role of NUDT14, a campaign was launched at the CMD to design a selective chemical probe for NUDT14 that would be useful in cellular function studies. Following previous literature examples of kinase inhibitor repurposing, a screen of a small library of known kinase inhibitors revealed somewhat surprisingly that ibrutinib (an FDA-approved Bruton's tyrosine kinase antagonist used in the treatment of B-cell cancers) was a dual inhibitor of NUDT14 and NUDT5.[375,381] An SAR campaign by Balikci *et al.* lead to the discovery of the dual antagonist **6.1** (Figure 6.1 panel A) with potency below 300 nM for both NUDT5 and NUDT14.

Analysis of the X-ray crystal structures of **6.1** bound to both NUDT5 and NUDT14 (Figure 6.1 panel A) revealed that whereas the π-stacking interactions of the heterocyclic core

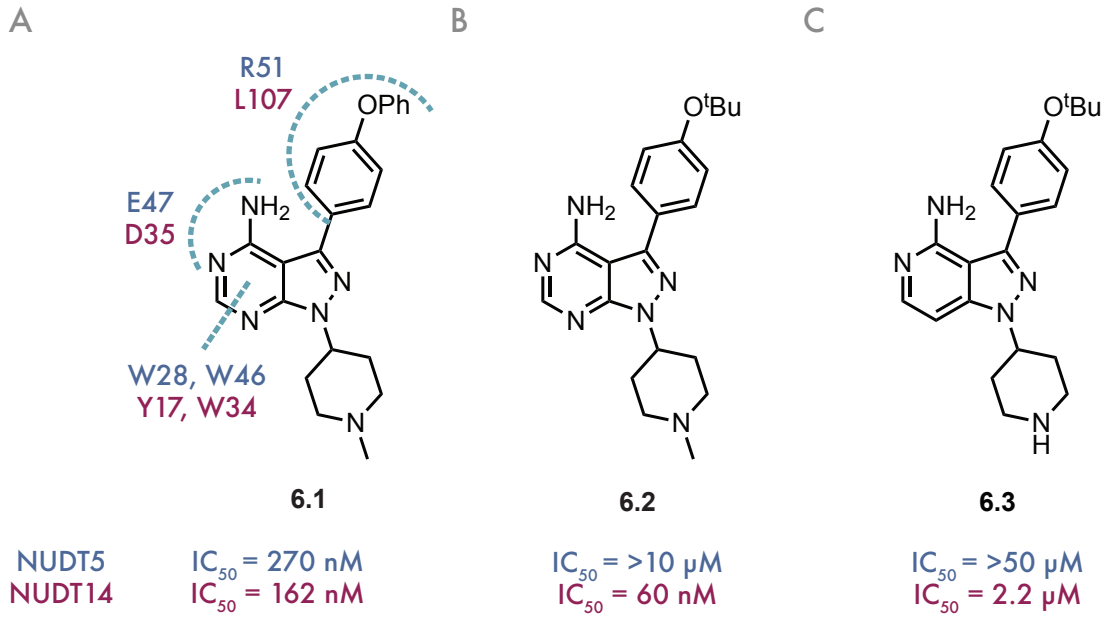


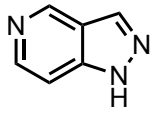
Figure 6.1. A | The imatinib-derived NUDT5-NUDT14 dual antagonist discovered by Balikci *et al.*, and its key interactions with the NUDT5 and NUDT14 binding pocket. This figure is adapted (with permission) from that originally created by Dr Anne-Sophie Marques. **B** | The NUDT14-selective chemical probe discovered by Marques *et al.*[a] **C** | The structure and potencies of **6.3**, designed using HCIE to investigate the effect of removing the pyrimidine nitrogen.

and the H-bond interactions with the upper portion of the aminopyrimidine ring are the same between the two proteins, the interactions with the 3-phenoxyphenyl moiety differ. Marques *et al.* therefore decided to focus SAR efforts on modifying the C3 position to improve selectivity. Extensive SAR analysis revealed that the 3-phenyl *tert*-butoxy sidechain of **6.2** (Figure 6.1 panel B) gave the best selectivity over NUDT5, and improved on the potency of **6.1**.

Wishing to explore the effect of bioisosteric modification of the heterocyclic core on the selectivity and potency, specifically with regards to exploring how adjusting its electronics affected the relative strength of $\pi - \pi$ stacking interactions between NUDT5 and NUDT14, a search of VEHICLe with version one of HCIE (the ShaEP implementation described in

[a]Manuscript in preparation at time of writing.

Chapter 2) was performed by Tom Dyer.[b] The results were inspected manually for synthetic feasibility, and to remove those which were missing exit-vectors in the desired positions. Due to the inclusion of PEB molecules[c] in the VEHICLe database, a very large number of the proposed candidates were rejected at this manual filtering stage. From the initial 24 867 heterocycles, a shortlist of seven was compiled. Although **pyrazolopyridine** was ranked in the bottom half of the VEHICLe database by HCIE, it was selected as a candidate based on its ready commercial availability, favourable exit vector geometry, and the ease with which its derivatives could be synthetically accessed. It also lacked the second pyrimidine hydrogen bond acceptor which, based on structural analysis of the binding pose (Figure 6.2), did not appear to contribute to any key interactions with the target but would alter the electronic properties of the ring. That pyrazolopyridine was not prioritised by HCIE despite sharing the same 6,5-bicyclic scaffold as the query ligand suggests that the key difference lies in the electrostatic potential of the core. Its ranking, while modest, was sufficient to warrant consideration, and the decision to prioritise it was informed by both practical considerations and a hypothesis-driven interest in probing electrostatic effects on NUDT14 selectivity.



Pyrazolopyridine

Interested to observe the effect that removing this nitrogen would have on the potency and selectivity, **6.3** (Figure 6.1 panel C) was designed, appending the most potent side chains to the new pyrazolopyridine core. At the time of compound design both the methylpiperidine (as in **6.2**) and the piperidine (as in **6.3**) sidechains were equipotent, however it was later

[b]Work done during a Part II project in the Brennan Group.

[c]Potentially explosive or bonkers; *vide supra*

found in a kinome screen that the piperidine analogues undesirably engaged RIPK2, which was avoided with the methylpiperidine, hence its appearance in the final compound **6.2**.

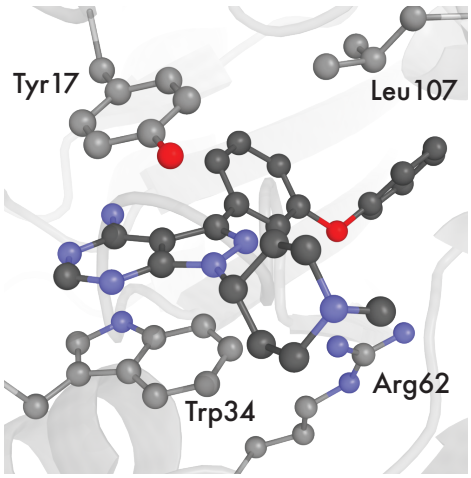


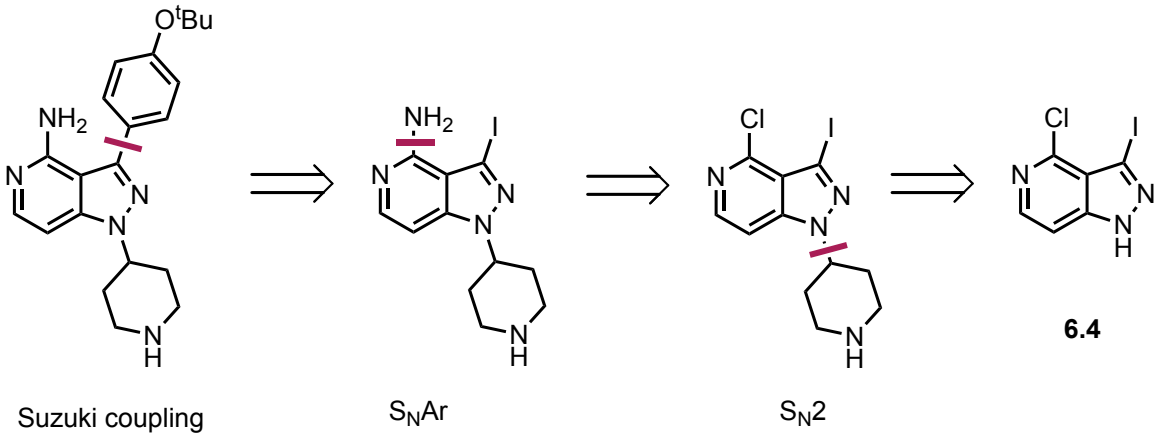
Figure 6.2. The structure of **6.1** bound to the active site of NUDT14 (PDB: 9GZV). The aromatic nitrogen γ to the amino group is facing towards the solvent, and does not appear to be forming any significant interactions with NUDT14 residues.

The chemical synthesis of **6.3** is outlined in Section 6.1.1 below. After successful synthesis, the compound was screened against NUDT5 and NUDT14 by Esra Balikci using AMP-Glo assays developed in the CMD, and the results of this screening are displayed in Figure 6.1 panel C.[375] Although the selectivity over NUDT5 was very encouraging, the deletion of the pyrimidine nitrogen does significantly reduce the NUDT14 potency from 60 nM for **6.2** to 2.2 µM in **6.3**. As the nitrogen did not appear to be forming any H-bonding interactions in the X-ray structure of the binding pocket, it is unlikely that a direct interaction with this nitrogen is the reason for the loss of potency. It is recognised that $\pi - \pi$ stacking interactions are generally more favourable between an electron-rich and an electron-deficient aromatic heterocycle, thus it is hypothesised that a reason for the drop in potency could be attributed to the amino-pyridine moiety being more electron rich than the amino-pyrimidine, and thus forming a less favourable π-stack.[87] It is also recognised that the preferred geometric arrangement of a $\pi - \pi$ stacking interaction is highly dependent on the interactions between

the dipoles of the rings involved.[264] Removing the pyrimidine nitrogen alters the dipole of the aromatic core, and thus is likely to alter the preferred geometry of the three-cycle $\pi - \pi$ stacking interaction in the binding pocket. If this alteration affects the ligand's binding pose, it could explain the drop in potency, either by forcing the ligand into a less favourable orientation within the binding pocket or by inducing conformational changes in the pocket itself to accommodate the ligand. Future work to determine this would involve attempting to find conditions where a crystal structure could be solved to investigate this.

6.1.1 Chemical Synthesis

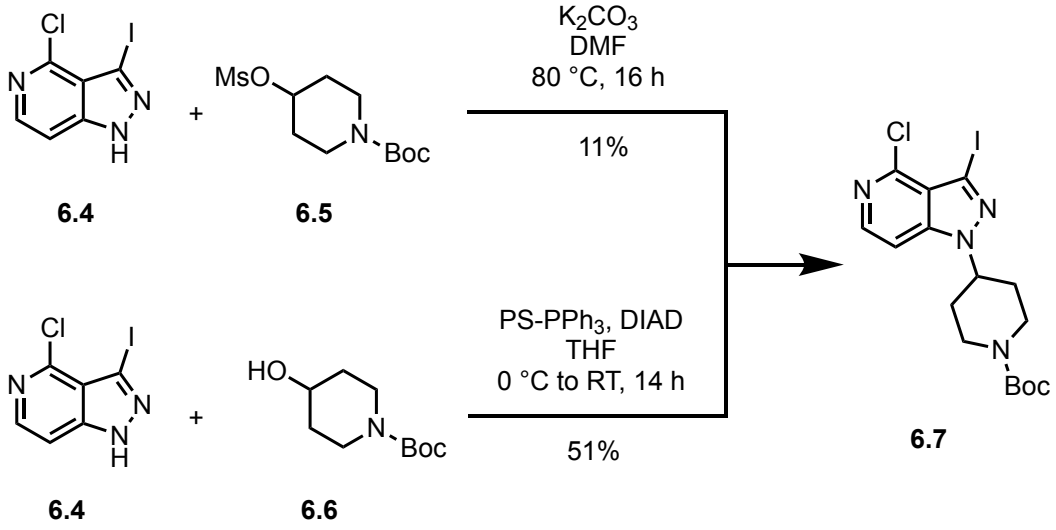
A retrosynthetic analysis of **6.3** is shown in Scheme 6.1. Taking advantage of the differing reactivity of aryl chlorides and aryl iodides towards S_NAr and Suzuki coupling, **6.3** could be taken back to commercially available 4-chloro-3-iodo pyrazolopyridine **6.4** in just three disconnections (excluding a single deprotection step outlined below).



Scheme 6.1: A retrosynthetic analysis of **6.3**.

The first route to **6.7** involved an S_NAr reaction between commercially available dihalopyrazolopyridine **6.4** and commercially available *N*-Boc protected piperidine-4-mesylate **6.5**, inspired by a similar route documented by Zhang *et al.* in 2016.[382] This is illustrated as the

upper route of Scheme 6.2. However this gave a disappointing 11% yield, which is attributed in part to the difficulty in separating the **6.5** starting material from the product. As this was the initial step, and the supply of **6.4** was limited due to cost, an alternative route with a higher yield and easier purification was sought.



Scheme 6.2: The synthesis of **6.7**. The top route is that initially explored using an S_NAr reaction, and the bottom the final route chosen, which makes use of a Mitsonobu reaction.

Inspired by a similar reaction documented by Sato *et al.*, a Mitsunobu reaction (the lower route of Scheme 6.2) was trialled for the formation of the pyrazole-piperidine N-C bond with commercially available alcohol **6.6**.[383] Pleasingly this resulted in a significantly improved yield of 51%, and the starting material **6.6** had a sufficiently different retention factor to the product that separation was facile by flash column chromatography.

The subsequent steps to synthesise **6.3** are shown in Scheme 6.3. It was decided to install the 4-amino group before any cross-coupling reactions to avoid issues of chemoselectivity, and thus an S_NAr reaction with 2,4-dimethoxylbenzylamine as an amine nucleophile selectively afforded **6.8** in 75% yield.[384] A global deprotection with TFA gave both the free piperazine and the 4-amino group in 88% yield. As the 3-iodo centre was highly activated, and to avoid any issues with unwanted Buchwald-Hartwig-type amination, $Pd(dppf)Cl_2$ was used

as a catalyst with a more restrained reactivity profile. After purification by preparative HPLC, this afforded the desired product **6.3** in low 7% yield. Despite the poor conversion of this reaction, sufficient product was isolated for biological testing, and thus due to time constraints the reaction was not further optimised.

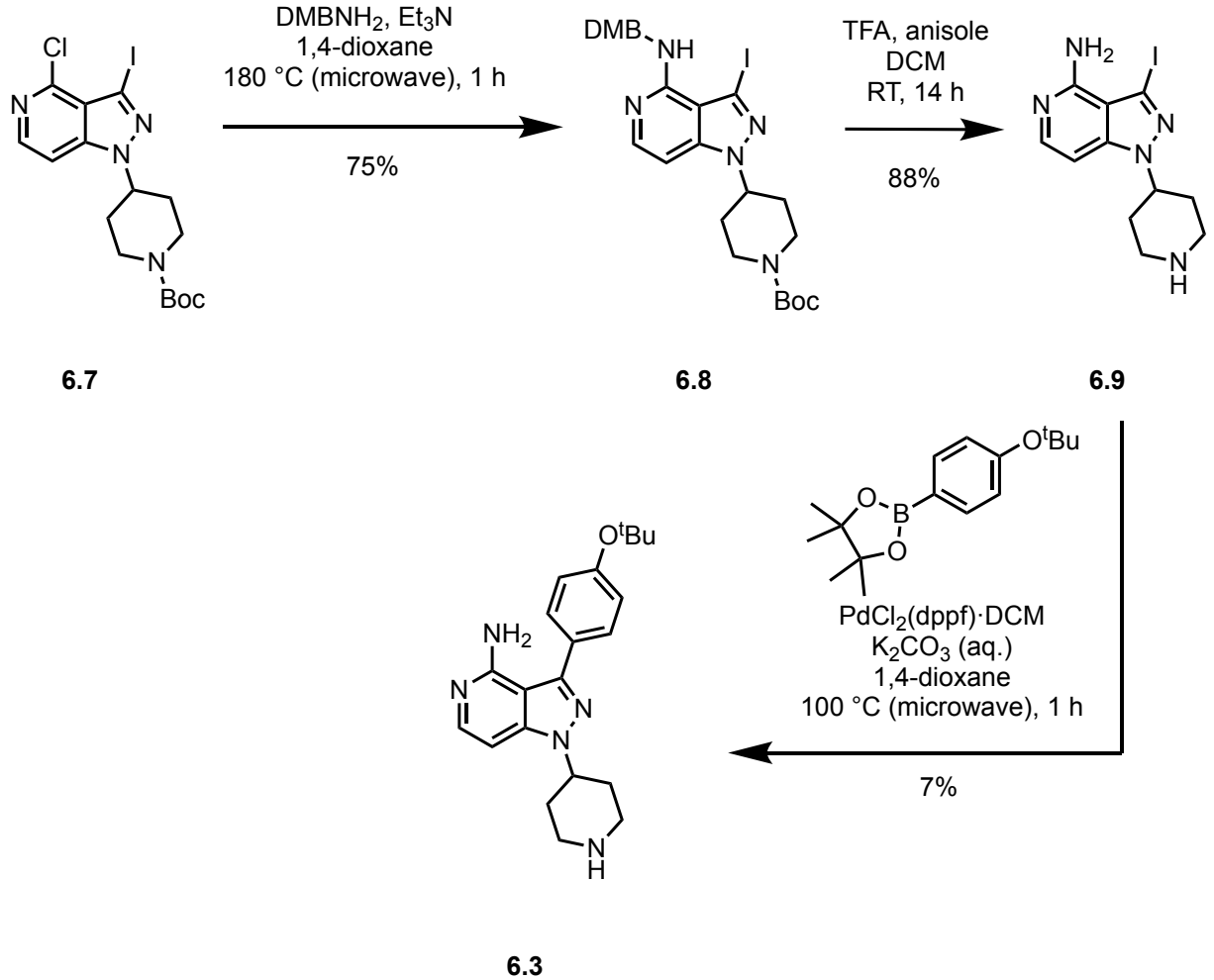
6.1.2 Conclusions and Future Work

This section has described the use of HCIE for designing a bioisosteric analogue of a potent ligand for NUDT14 discovered in the CMD, and the synthetic route used to access it. Although the desired selectivity over NUDT5 was maintained, the potency was reduced from 60 nM to 2.2 µM. Several reasons for this are proposed, based on a modulation of the energetics and geometry of the $\pi - \pi$ stacking interaction observed in the binding pocket by altering the electronics of the aromatic core. Although the potency was reduced, there was still biochemical evidence of binding to NUDT14, and as no further optimisation of the sidechains around this core or pharmacokinetic profiling of the compound was carried out, it is still possible that this substitution may lead to an improvement in other desirable properties.

The results in this section also serve to highlight a key limitation of ligand-based approaches to bioisostere discovery: the tolerance for deviations in shape or electrostatics can vary significantly between targets. It is not possible to characterise these tolerances without conducting experimental profiling of the binding site, which involves observing the effects of systematically varying the electronics and sterics of ligands on the binding affinity. That the pyrazolopyridine core selected here was ranked in the bottom half of the HCIE results despite having the exact same 6,5-bicyclic architecture as the query ligand suggests a significant ESP difference between the two cores. The observed drop in potency is therefore consistent

with a binding site that is sensitive to electrostatic changes. The deliberate selection of a lower-scoring HCIE molecule allowed the probing of this sensitivity directly, and contributed towards a better understanding of the NUDT14 binding site.

This compound was designed before the second generation HCIE software (as documented in Chapter 4) was developed, and thus future work will involve using the latest iteration of HCIE to again search for the heterocyclic core of **6.2** with the new search functionality and the expanded MoBiVic library, as well as exploring structurally the reasons for the drop in potency.

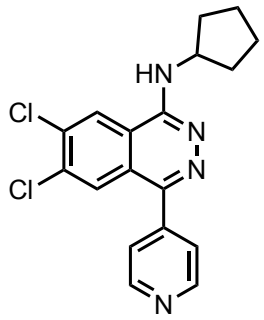


Scheme 6.3: The synthesis of **6.3** from **6.7**, affording **6.3** in overall 3% yield over 4 steps.

6.2 Small-Molecule Inhibitors of SUV4-20

Friedreich's Ataxia is a rare neurodegenerative genetic disorder and is the most common hereditary ataxia, with an estimated prevalence of 1 in 50 000 and a European carrier frequency of 1 in 120.[385] The primary phenotype is an ataxia (lack of muscular coordination and control) of all four limbs, often occurring before the patient is 20 years old and degenerating with age, leaving patients confined to a wheelchair and with a mean life expectancy of just 37 years.[386,387] It is an autosomal recessive disorder caused by a GAA trinucleotide repeat expansion in the first intron of the *FXN* gene encoding frataxin, a mitochondrial protein playing a critical role in iron homeostasis.[385,388,389] Whilst intronic GAA repeats are not abnormal, they are much more numerous in Friedreich's ataxia-associated alleles ($\sim$ 40 in wild type, 600-900 in affected individuals), and lead to a partial silencing of frataxin expression by the formation of unusual DNA structures and induction of epigenetic changes, particularly the increase in repressive chromatin marks (for example H4K20 trimethylation). The number of repeats has been directly correlated with the severity and age of disease onset.[386]

SUV4-20 H1 and SUV4-20 H2 (collectively SUV4-20) are homologous histone methyltransferases responsible for di- and trimethylation of histone H4 at lysine 20 (H4K20), an evolutionarily conserved modification essential for chromatin structure and genomic stability.[390–392] Studies by Vilema-Enríquez *et al.* found that inhibiting SUV4-20 raised frataxin expression in a human luciferase reporter cell model, and also in primary Friedreich's ataxia patient-derived cells.[389] An HTS campaign by Bromberg *et al.* identified **A-196** as a potent and selective probe for the SUV4-20 enzymes, with Vilema *et al.* subsequently showing that **A-196** increases frataxin expression by a factor of 1.5, a clinically-significant increase for preventing further disease degradation in patients.[385,389,393]



A-196

SUV4-20 H1 IC_{50}	25 nM
SUV4-20 H2 IC_{50}	144 nM
FXN Expression EC_{50}	5.2 µM
Methyltransferase selectivity	50×
HLM[a] Cl_{int}	191 µL min^{-1} mg^{-1}
HLM[a] $t_{\frac{1}{2}}$	7.3 min
MLM[b] Cl_{int}	239 µL min^{-1} mg^{-1}
MLM[b] $t_{\frac{1}{2}}$	5.8 min

Figure 6.3. The structure of **A-196**, its potencies against the SUV4-20 enzymes, and its metabolic profile.[394] IC_{50} measurements against SUV4-20 enzymes were determined by a scintillation proximity assay with radiolabelled SAM. [a] Human liver microsomes. [b] Mouse liver microsomes.

Although **A-196**'s potency, selectivity, and effect on frataxin expression are excellent, its metabolic instability (specifically its short half-life and high rate of intrinsic clearance; see Figure 6.3) precludes its *in vivo* use. Dr Rob Quinlan designed and synthesised 140 analogue structures of **A-196** to improve its metabolic stability.[394] These were each assayed for SUV4-20 H1 inhibition in an MTase-Glo™ bioluminescence-based assay at four separate concentrations, and their effects on the thermal stability of SUV4-20 H1 also measured in a nanoDSF (differential scanning fluorimetry) assay.[395,396] Based on these results, and using the crystal structure of **A-196** bound to SUV4-20 H1 (PDB: 5CPR), Xinyu Chen built a docking & scoring protocol and a quantitative structure-activity relationship (QSAR) model for small-molecule ligand binding to SUV4-20 H1.[393,397d] Integrating both of these models into a virtual screening pathway, compounds from the Enamine CNS library, a selection of the BioAscent library[e], and a subset of the full Enamine screening library were screened against SUV4-20, with those showing positive results further filtered for desirable physicochemical properties before being visually inspected. From an initial screening pool of $> 100\,000$ com-

[d]Work done in the Brennan Group as part of a Part II project.

[e]Those with Enamine as a supplier were selected.

pounds, 168 were purchased and screened by MTase-Glo™ and nanoDSF against SUV4-20 H1.[f] Seven compounds displayed methyltransferase activity, and six of these are displayed in Figure 6.4.[g]

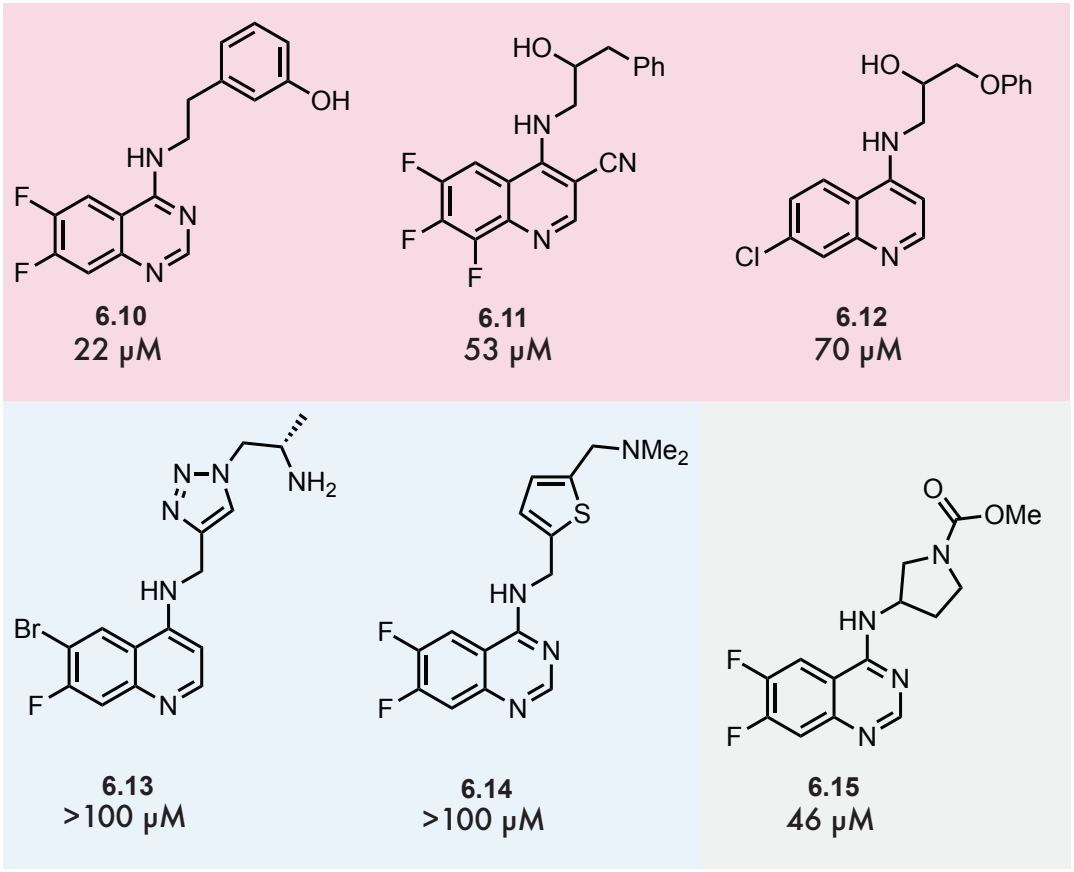


Figure 6.4. Six of the structures from the virtual screening campaign by Xinyu Chen in the Brennan Group that displayed experimental methyltransferase inhibition. The IC_{50} values, as measured by the MTase-Glo™ assay performed by Crelux GmbH., are shown for each compound. The colours illustrate the classification of the compounds as outlined below.

From these structures and results, an analysis of the docked structures in the SUV4-20 binding pocket (not shown) allowed a crude separation of these compounds into two classes:

[f]This screening was performed at Crelux GmbH.

[g]The seventh compound was based on a separate 5,6-bicyclic heteroaromatic core, was the least potent, and did not fit into this project, and so was not considered.

1. Those with flexible, 2-4 carbon linkers to aromatic rings (highlighted in pink in Figure 6.4). These have polar hydroxyl groups either on the linker (**6.12** and **6.11**) or on the aromatic ring (**6.10**). It was suggested based on the docked structures that these could form an additional H-bond with Val-252 in the binding pocket.

2. Those with a five-membered aromatic ring bonded to a basic methyl amine sidechain (**6.13** and **6.14**; highlighted in blue). It was suggested based on docked structures that these could form an additional $\pi - \pi$ stacking interaction with Tyr-307 in the binding pocket.

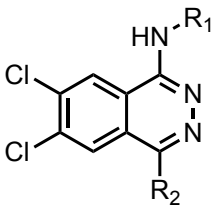
6.15 (highlighted in grey in Figure 6.4) did not appear to fit neatly into either of these two classes of compound, instead bearing structural similarities to both. All of these compounds display far weaker potencies than **A-196**, and it was hypothesised that this was due to the different cores and lack of 4-pyridyl side chains when compared with **A-196**. The 6,7-dichlorophthalazine core of **A-196** was not available in the Enamine compound catalogue, and thus to investigate the potency of the sidechains discovered in this virtual screen, compounds were designed that fused these sidechains with the dichlorophthalazine core of **A-196**, and either the original 4-pyridyl or an *N*-methyl pyrazole sidechain discovered as part of Dr Rob Quinlan's initial analogue investigations.

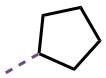
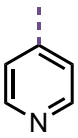
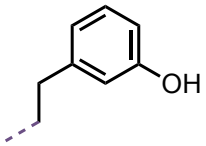
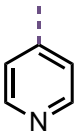
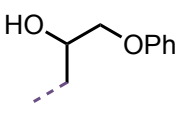
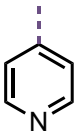
Recognising that the dichlorophthalazine core of **A-196** had a common exit-vector geometry, it was hoped that the implementation of HCIE described in Chapter 4 could be used to identify bioisosteres of this core, potentially with improved pharmacokinetic properties. It was therefore decided that the sidechains proposed in Figure 6.4 would be synthesised on the dichlorophthalazine core initially. The most potent sidechains would then be selected and fused to a selection of high-scoring cores from a HCIE search, and these synthesised,

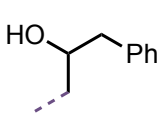
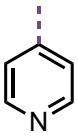
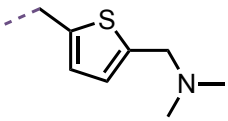
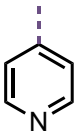
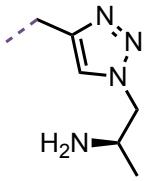
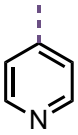
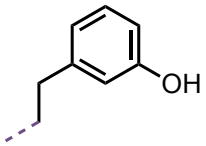
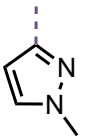
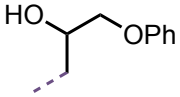
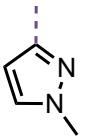
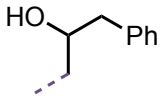
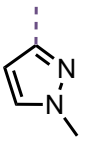
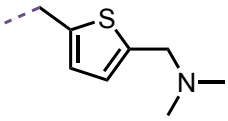
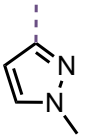
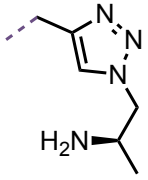
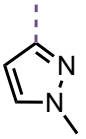
assayed, and metabolically characterised to observe the effects of replacing the core on the pharmacokinetic profile of the compounds.

The initial compounds, with the sidechains fused to the dichlorophthalazine core, are shown in Table 6.1, along with their experimental IC_{50}s obtained in a biochemical assay. It was initially decided to focus on the compounds that sat clearly in one of the two classes described above, thus the pyrollidine methoxyester of **6.15** was not included in these initial compound designs. The chemical synthesis of the selected compounds is discussed in Section 6.2.1.

Table 6.1. The compounds designed based on the results of the virtual screen and Dr Rob Quinlan's analogue syntheses and biological characterisation. The IC_{50} results are for SUV4-20 H1 and are from an MTase-Glo™ assay conducted by WuXi AppTec. **A-196**'s potency was assayed as a reference. [a] Compound unsynthesised at the time of writing.



Compound	R_1	R_2	Inhibition	IC_{50} (µM)
A-196			Full	0.60
6.16			Partial	0.63
6.17			Partial	2.83

Compound	R_1	R_2	Inhibition	IC_{50} (µM)
6.18			None	—
6.19[a]			N.D.	N.D.
6.20[a]			N.D.	N.D.
6.21			Full	1.20
6.22			Partial	2.90
6.23			Partial	3.69
6.24[a]			N.D.	N.D.
6.25[a]			N.D.	N.D.

Of the compounds that were synthesised initially, all except **6.18** showed clear dose-responses with significantly improved potency over their analogues from the virtual screen. This is

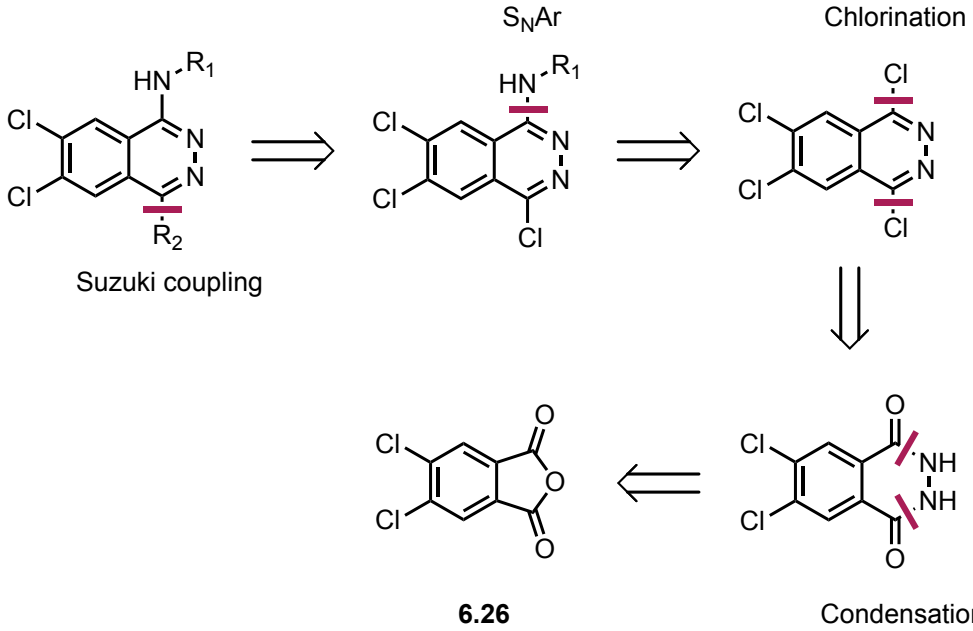
pleasing as it demonstrates the effectiveness of the $\mathbf{R_2}$ group and 6,7-dichlorophthalazine core at improving binding through the additional interactions in the binding pocket, but it also validates the virtual screening approach for identifying potent side-chains at a lower synthetic cost than a traditional, synthesis-based SAR campaign. Interestingly, the order of binding affinity of the side-chains established in Figure 6.4 is maintained in these elaborated compounds, with the 3-phenol sidechain of **6.10** showing the highest potencies and the 3-phenylpropan-2-ol sidechain the lowest across both $\mathbf{R_2}$ analogues.

Full inhibition was only observed for **A-196** and **6.21**, which showed an IC_{50} of 1.20 µM; a factor of two away from that observed for the reference **A-196** of 0.6 µM. This result is encouraging, as the most potent compounds identified for this scaffold as part of the previous SAR activity had IC_{50} values greater than 10× that of **A-196**.

There is a large difference in the IC_{50} of the **A-196** reference compound as measured by this MTase-Glo assay and the radioactive scintillation assay used to measure the IC_{50} values reported in Figure 6.3. Differences in absolute value are to be expected when fundamentally different assays are used to measure a biological effect, however the magnitude of the difference here spans several orders of magnitude, and is a subject of ongoing investigation. Although it is advisable not to compare absolute potency values at this stage, the potencies of the compounds in Table 6.1 relative to the **A-196** potency measured in the same assay are nonetheless encouraging, and provide a platform for further elaboration and pharmacokinetic investigation.

6.2.1 Chemical Synthesis

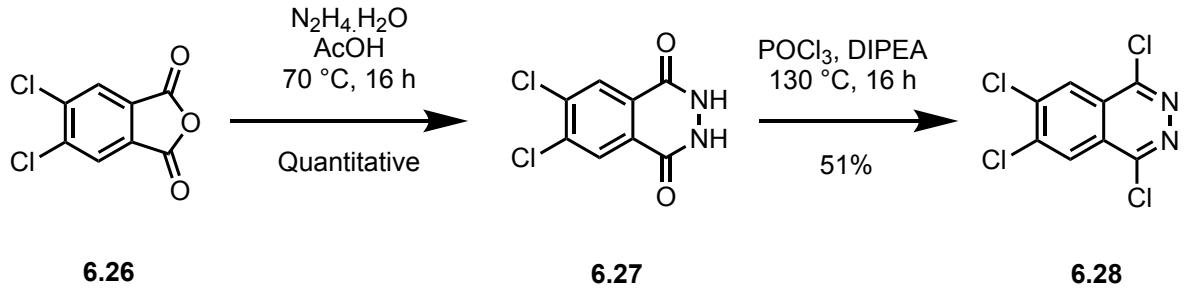
The syntheses of the compounds in Table 6.1 were based on a retrosynthetic analysis conducted by Dr Rob Quinlan in the Brennan Group, and illustrated in Scheme 6.4.[394] The Suzuki coupling was selected as the final step for a late-stage divergence strategy, with the common core and side chains functionalised with either a pyridyl or methylpyrazole substituent at the last stage of the route.



Scheme 6.4: The retrosynthetic analysis of the proposed SUV420 compounds, based on a route initially designed by Bromberg *et al.*[393]

Synthetic efforts began with the synthesis of the tetrachlorophthalazine core **6.28**, which was accessed via the commercially available 4,5-dichlorophthalic anhydride **6.26** using a route based on that proposed by Lanman *et al.* (Scheme 6.5).[398] A condensation reaction in acetic acid solvent with hydrazine monohydrate gave dihydrophthalazine-1,4-dione **6.27** in quantitative yield. After removal of the acetic acid solvent *in vacuo*, the resulting white powder was subject to a dehydrative chlorination with $POCl_3$ and DIPEA to give the desired

tetrachlorophthalazine **6.28** in 51% yield. Although the yield of the chlorination step was modest, this was isolated in sufficient multi-gram quantities for subsequent steps.



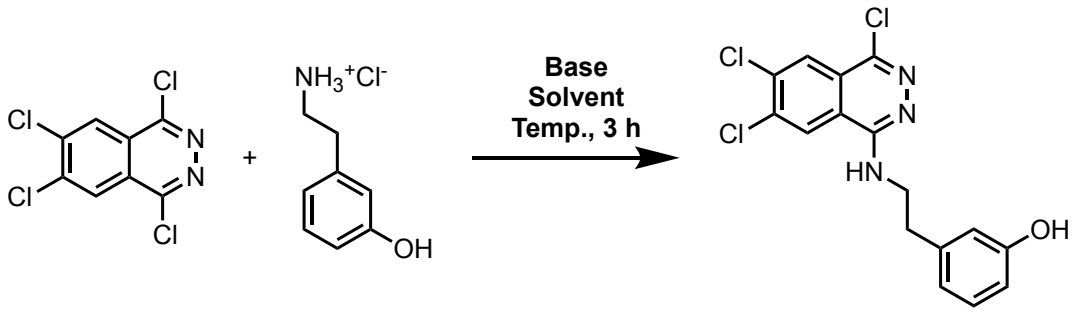
6.26 **6.27** **6.28**

Scheme 6.5: The synthesis of tetrachlorophthalazine **6.28** from dichlorophthalic anhydride.

The subsequent step involved installation of the secondary amine substituent by S_NAr reaction with tetrachlorophthalazine **6.28**. Previous analogues synthesised in the Brennan Group had made use of DMF and triethylamine base to perform this transformation, however attempts under these conditions with 3-hydroxyphenylethylamine hydrochloride **6.29** only returned starting material (Table 6.2 entry 1). A screen of conditions (Table 6.2) demonstrated that this transformation was unexpectedly sensitive to solvent and base, however IPA with DIPEA as a base gave a sufficient 74% yield after three hours at 90 °C, and thus these conditions were used for all further S_NAr reactions.

Initial efforts were focused on the compounds for which the amine sidechain starting materials were commercially available (**6.16, 6.17, 6.18, 6.21, 6.22, 6.23**). The conditions described above afforded the *N*-aryl products **6.30**, **6.31**, and **6.32** in sufficient yields for the Suzuki coupling (Table 6.3, upper panel). Recognising that the desired 1-chloro position was highly activated due to its proximity to the electron-withdrawing phthalazine nitrogens, and wishing to avoid side-reactivity at the unactivated chlorides in the 6 and 7 positions, all couplings were initially attempted with $Pd(dppf)Cl_2$ catalyst and aqueous sodium carbonate in 1,4-dioxane. Although these conditions gave satisfactory isolated yields for the compounds

Table 6.2. A screen of conditions for the nucleophilic aromatic substitution of **6.29** with **TCPh**. *Reagents and conditions:* **TCPh** (0.08 mmol), amine hydrochloride salt (1 eq.), base (3 eq.), solvent (1 mL). [a]Isolated yield. [b]As observed by LCMS of crude reaction mixture.



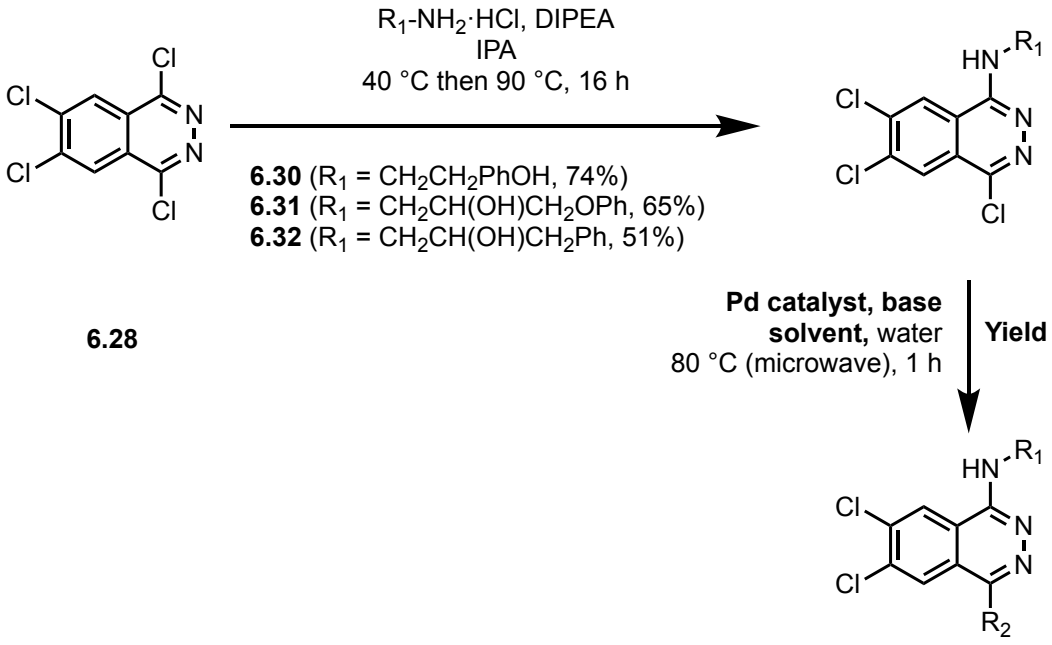
6.28 **6.29** **6.30**

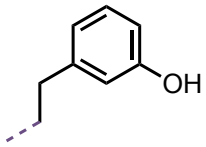
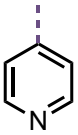
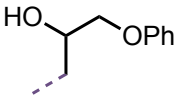
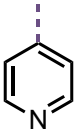
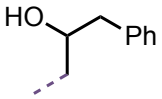
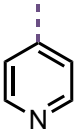
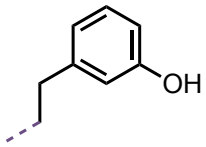
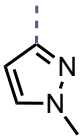
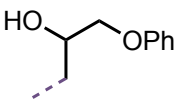
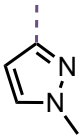
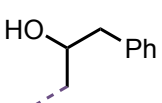
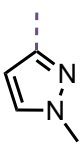
Entry	Solvent	Base	Temperature (°C)	Yield
1	DMF	Et_3N	70	0%[a]
2	IPA	K_2CO_3	70	0%[a]
3	IPA	DIPEA	90	74%[a]
4	DMA	DIPEA	70	0%[b]
5	DMA	Cs_2CO_3	70	0%[b]
6	DMA	Na_2CO_3	70	0%[b]
7	DMA	Et_3N	70	0%[a]

coupled to the methylpyrazole substituent, those coupled to 4-pyridyl were difficult to purify as diphenylphosphineferrocene oxide eluted with the product under all attempted purification conditions. Preparative HPLC was able to isolate pure **6.18** in 9% yield, which provided sufficient product for biological testing, however **6.16** and **6.17** could not be isolated in sufficient purity by this method.

Further coupling conditions for the synthesis of **6.16** and **6.17** were screened. An attempt at accessing **6.16** with XPhos Pd G2 catalyst and sodium carbonate base in 1,4-dioxane and water at 40 °C (not shown) gave a mixture of 4,6,7-trisubstituted phthalazine and starting material. Reaction with tetrakistriphenylphosphine palladium and aqueous potassium carbonate in THF gave significant quantities of triphenylphosphine oxide, but **6.17** was successfully purified in 16% yield by preparative HPLC. The use of polymer-supported

Table 6.3. The syntheses of **6.16-6.18** by S_NAr and Suzuki Coupling. [a]Polymer-supported catalyst.



Compound	R_1	R_2	Catalyst	Base	Solvent	Yield
6.16			$Pd(PPh_3)_4$[a]	K_2CO_3	THF	8%
6.17			$Pd(PPh_3)_4$	K_2CO_3	THF	16%
6.18			$Pd(dppf)Cl_2$	Na_2CO_3	dioxane	9%
6.21			$Pd(dppf)Cl_2$	Na_2CO_3	dioxane	23%
6.22			$Pd(dppf)Cl_2$	Na_2CO_3	dioxane	58%
6.23			$Pd(dppf)Cl_2$	Na_2CO_3	dioxane	65%

triphenylphosphine palladium under the same conditions enabled successful purification of **6.16**, albeit with a low 8% yield.

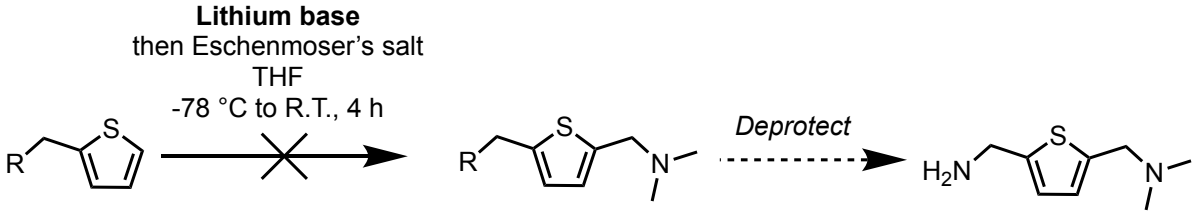
6.2.1.1 Efforts towards 6.19 and 6.24

The proposed routes to access **6.19** and **6.24** were based on those used to access **6.16 - 6.18**, however as the diaminomethyl thiophene **6.36** was not available commercially, a route to access this building block based on an ortholithiation strategy was explored (Table 6.4, upper panel). Recognising that **6.36** resembles a reduced Vilsmeier-Haack intermediate, it was hypothesised that trapping an ortholithiated aminomethyl thiophene with Eschenmoser's salt might yield the desired product.

Attempts to access **6.36** without addition of organolithium showed no signs of product formation (Table 6.4 entry 1), and attempts at ortholithiation with *n*-BuLi and LDA (both commercial and freshly prepared LDA) gave no reaction. Attempts to trap the lithiated product with D_2O or MeOD also showed no evidence of the formation of ortholithiated product. Hypothesising that the free amine group might be interfering with the organolithium or quenching the lithiated intermediate, phthalimide protected and STABASE protected amine analogues were prepared and ortholithiations attempted (entries 4 and 5). Unfortunately these reactions also showed no evidence of desired product formation, and attempted trapping of the lithiated intermediate with deuterated water were inconclusive. Further efforts to access this intermediate are ongoing, and will involve screening other, more basic organolithiums including *sec*-BuLi and *tert*-BuLi.

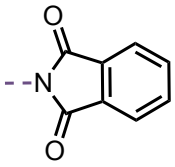
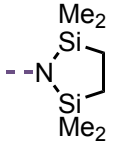
In parallel with the ortholithiation attempts, a strategy involving the reductive amination of phthalimide protected **6.34** was also investigated (Scheme 6.6). Following a route taken from

Table 6.4. Results of attempts to install an *N,N*-dimethylmethylamine sidechain via an ortho-lithiation, with Eschenmoser's salt as a trapping electrophile. *Reagents and conditions:* 2-thiophenemethylamine (0.8 mmol) was added to a solution of **lithium base** (2 eq.) in dry THF (4 mL) at -78 °C and stirred for one hour, after which Eschenmoser's salt (1.5 eq.) was added, and the mixture warmed to room temperature and stirred for a further three hours. N.R. = no reaction. [a] Reaction attempted in both THF and MeCN.

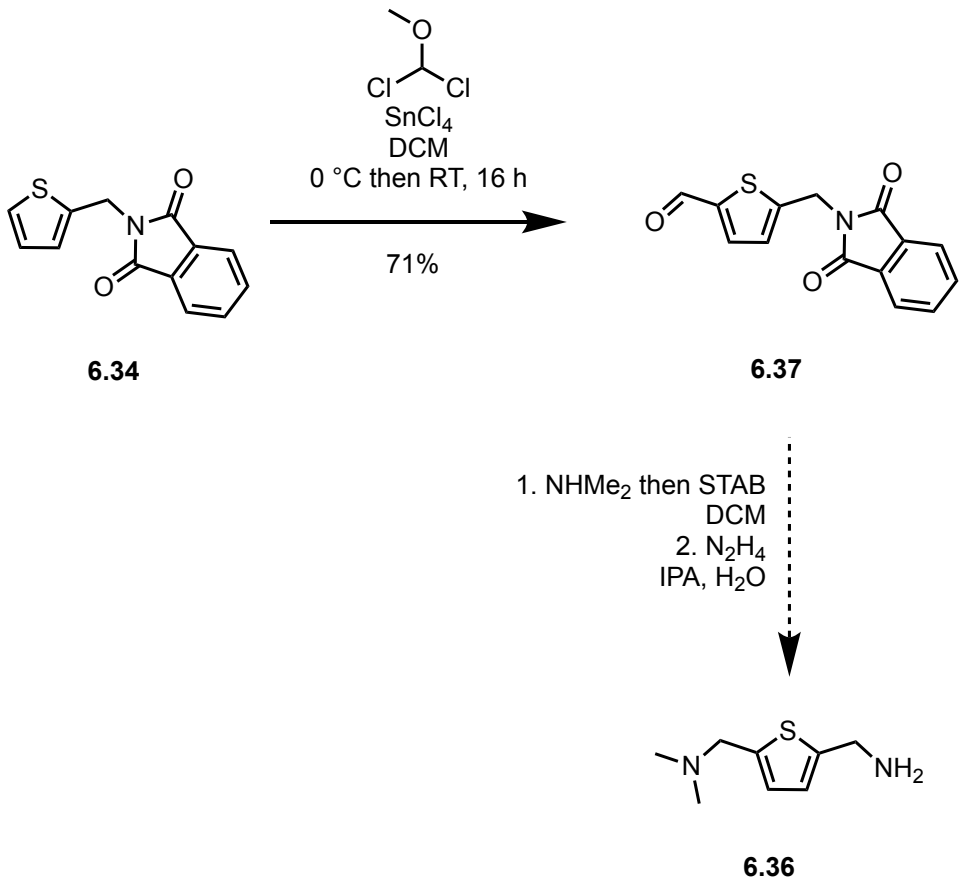


6.33 (R = NH_2)
6.34 (R = $N(CO)_2C_6H_4$)
6.35 (R = $N(SiMe_2)_2C_2H_4$)

6.36

Entry	R	Base	Outcome
1[a]	--NH_2	—	N.R.
2	--NH_2	*n*-BuLi	N.R.
3	--NH_2	LDA	N.R.
4		LDA	N.R.
5		*n*-BuLi	N.R.

a 1999 Glaxo Wellcome Inc. patent, a Rieche formylation on phthalimide **6.34** afforded the desired aldehyde **6.37** in 71% yield.[399] Investigation of the feasibility of reductive amination with dimethylamine followed by phthalimide deprotection with hydrazine is a subject of ongoing work on this project.



Scheme 6.6: Synthesis of **6.36** by Rieche formylation, and proposed reductive amination and phthalimide deprotection.

6.2.2 Future Work and Conclusions

Regrettably time constraints meant that it was not possible to finish the synthesis of the initial compounds, or to include any cores proposed by a HCIE search, therefore continuing work on this project will focus on two key areas:

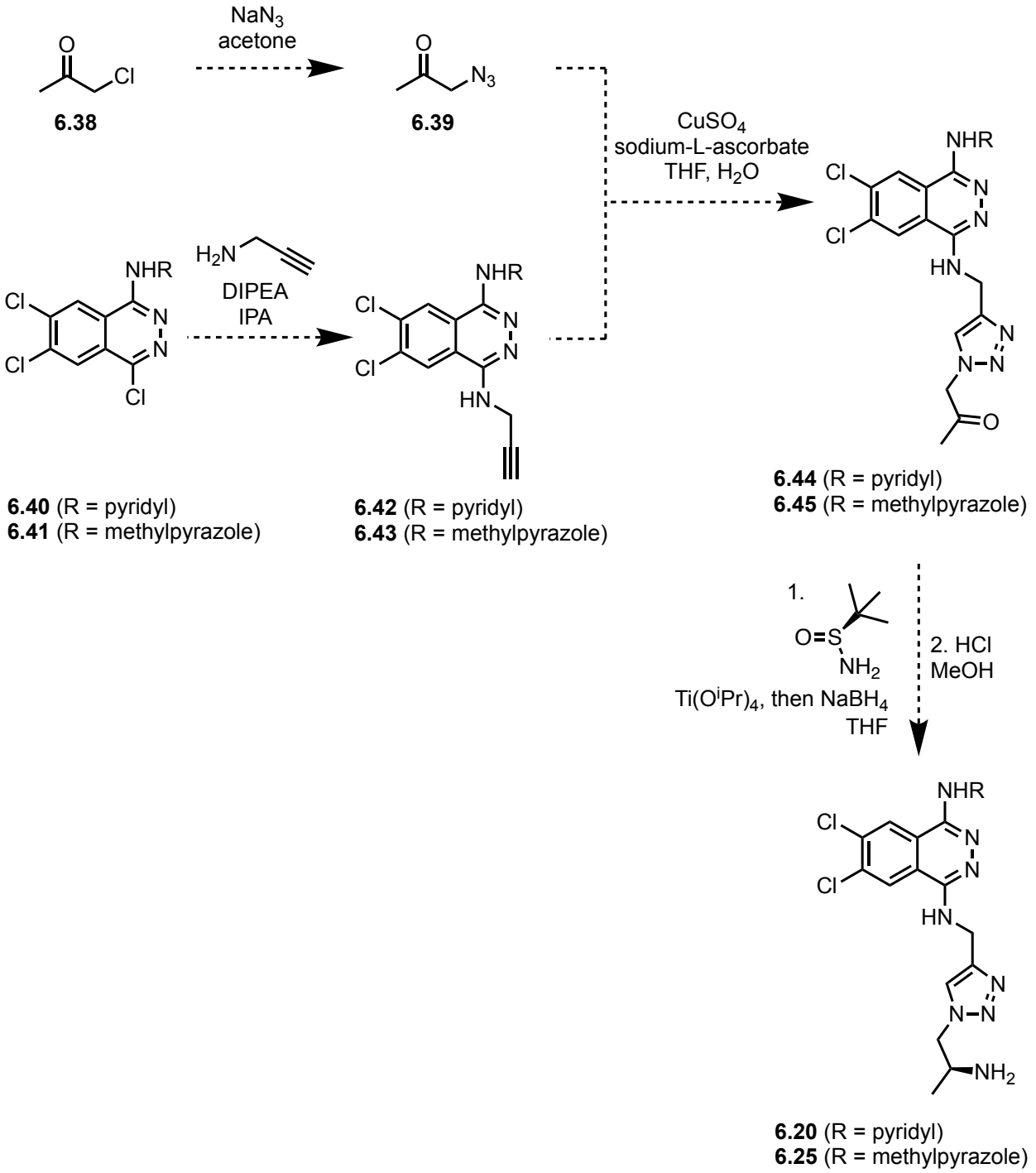
1. Completion of the synthesis of the compounds in Table 6.1, and their biological characterisation.

2. Design, synthetic exploration, and biological characterisation of bioisosteric replacements of the 6,7-dichlorophthalazine core using HCIE.

As previously discussed, the discrepancies between the IC_{50} values for **A-196** as measured by scintillation and MTase-Glo assays are a subject of ongoing investigation, and immediate future work will involve characterising the ligands in Table 6.1 in the scintillation assay so that their potencies can be directly compared to the ligands discovered in Dr Rob Quinlan's SAR investigation. Furthermore, as a key objective of this project is to improve the metabolic stability of the SUV4-20 inhibitors, the metabolic profile of the ligands synthesised here will need to be characterised to determine the effect of the novel sidechains on the half-life and clearance of the ligands.

Completing the exploration of the ortholithiation route towards **6.36** through further screening of organolithium bases and experimental conditions is currently ongoing, with the reductive amination and deprotection of **6.37** as an orthogonal route to **6.36** also being explored. The successful preparation of **6.36** would then allow the preparation of **6.19** and **6.24** following the synthetic scheme outlined in Table 6.3.

An asymmetric synthetic route towards **6.20** and **6.25** was designed (Scheme 6.7), and further future work will involve its implementation and optimisation to obtain the enantiomerically-pure desired compounds. This route involves an early-stage Suzuki coupling to avoid issues of chemoselectivity presented by the ketone of **MH094** or the primary amine **6.20/6.25**, followed by an S_NAr reaction with propargylamine to install an alkyne handle. A copper-promoted Huisgen cycloaddition (click reaction) with 1-azidopropan-2-one (prepared from an S_N2 reaction of chloroacetone and sodium azide) to give the triazine, and then an asymmetric Ti(IV) catalysed reductive amination with Ellman's sulfinamide would yield the desired amine in good enantiomeric excess. These compounds will then be assayed for SUV420 inhibition.



Scheme 6.7: The proposed asymmetric synthesis of **6.20** and **6.25** by Huisgen cycloaddition and Ellman's sulfinamide.

The second area of future research involves further exploration of the central core of these molecules, making use of HCIE to identify bioisosteric cores to the 6,7-dichlorophthalazine used in these molecules. A HCIE search of the 1,4-disubstituted-6,7-dichlorophthalazine was conducted, and the top 10 returned molecules (excluding the query itself) is shown in Figure 6.5. All of the proposed molecules are sterically similar to the original core, and all of the proposed substituents are in the same 6,7 positions, however it is interesting to note the range of both electron donating and electron withdrawing substituents, and H-bond donors and acceptors proposed by HCIE. **S645003** is particularly interesting, as previous work by Dr Rob Quinlan has shown that the 6,7-dimethoxy core is not well tolerated by SUV420, which was attributed to the need for the dimethoxy substituents to adopt a sterically unfavourable gauche conformation, however the single methoxy substituent has not been synthesised or assayed previously.[394] Methoxy groups are known bioisosteres of chloro substituents (a SwissBioisostere search shows that chloro groups have been substituted for methoxy groups 12 972 times, improving or retaining potency in 9740 of these cases), and lower the lipophilicity of the molecule (a key consideration in the previous SUV420 campaign to reduce metabolic clearance) whilst maintaining hydrophobic interactions with the SUV420 binding pocket.[400] Thus it will be interesting to observe the effect of **S645003** on the inhibition of SUV420, and the pharmacokinetic profiling of the resulting ligands relative to the original dichloro core.

Docking studies (lead by Dr Laura Ortega-Varga) are currently being pursued, and the results of these will be used to select cores for further synthesis and biological evaluation.

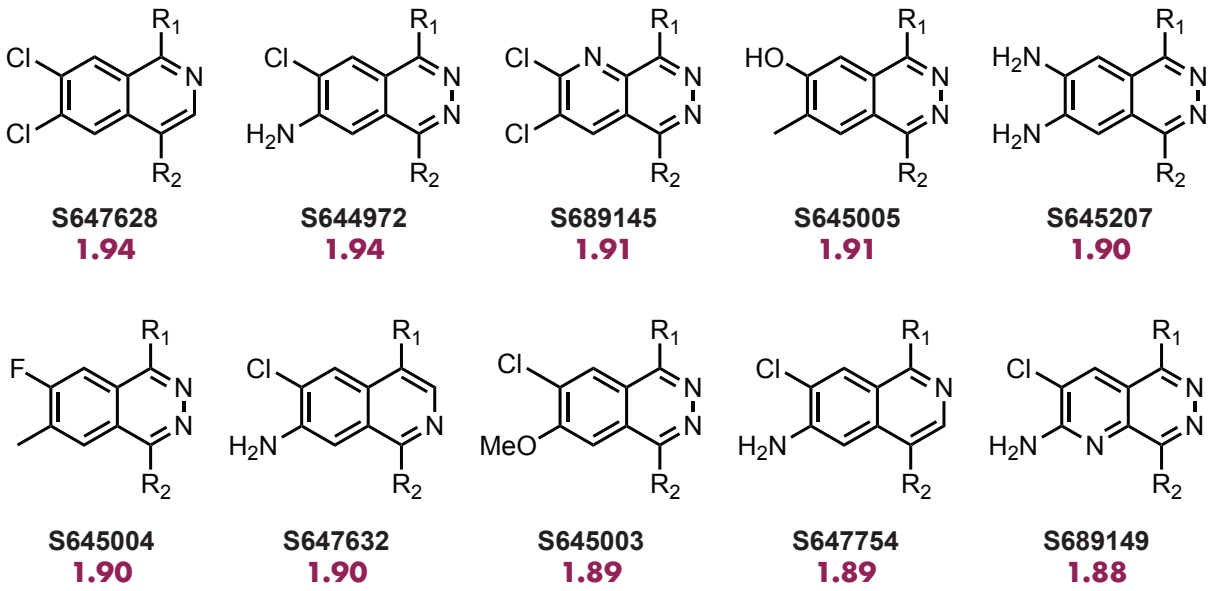


Figure 6.5. The top 10 highest scoring results of the HCIE search for 1,4-disubstituted-6,7-dichlorophthalazine and their total scores. The query itself was returned as the highest scoring molecule, and is not shown here.

7 Thesis Conclusions

Aromatic heterocycles make frequent appearances in small molecule drug discovery on account of the control they afford over the physicochemical, pharmacokinetic, and metabolic properties of a drug candidate. Their rigid geometry enables precise positioning of hydrogen-bonding functionalities to optimise target binding and selectivity, while their tunable electronics allow precise adjustment of physicochemical properties and metabolic stability. The properties of aromatic heterocycles that favour their inclusion in small molecule drugs also make them particularly well suited for use as bioisosteres. Importantly, even small modifications such as constitutional changes or ring fusions can lead to meaningful changes in physicochemical properties without compromising the molecule's ability to engage its target. As such, aromatic heterocycles are frequently deployed in scaffold-hopping or functional group replacement strategies aimed at improving potency, selectivity, or ADMET properties. Given the pharmacological significance of aromatic heterocyclic bioisosterism, there is a need for systematic methods to identify and prioritise aromatic heterocyclic bioisosteres, particularly as their combinatorial space far exceeds what can be explored empirically.

Taking Alfred Burger's 1991 definition of bioisosterism (*compounds or groups that possess near-equal molecular shapes ... approximately the same distribution of electrons ... and*

biological properties that are related to each other) as a foundation, this thesis describes the conception, implementation, and evaluation of computational tools (principally the Heterocycle Isostere Explorer; HCIE) for identifying novel aromatic heterocyclic bioisosteric replacements of medicinally-relevant heterocycles.

Chapter 2 described the development and initial testing of a first-generation implementation of HCIE, which employed the VEHICLe database of aromatic heterocycles as a searchable space of potential bioisosteres. A shape and electrostatics-based comparison of query and database molecules was carried out using the ShaEP algorithm developed by Vainio *et al.*, with geometries and partial charges derived from xTB semi-empirical DFT calculations. Although this implementation showed a promising correlation between bioactivity and similarity scores when benchmarked against inhibitors of SARS-CoV-2 MPro, particularly when higher-level DFT-derived partial charges were employed, the dataset used in benchmarking ultimately proved to have limitations. These included the variability of inactive compounds and the flexibility of the MPro binding site, which made classification of active and inactive compounds challenging. Furthermore, limitations in the methodology were identified, particularly regarding the appropriateness of the alignment and scoring algorithm in the SHaEP implementation for measuring similarity in simple aromatic heterocycles, and the restricted scope of the VEHICLe library which lacks functionalised heterocycles. These findings informed the design of the MoBiVic database (Chapter 3) of functionalised aromatic heterocycles, and the development of a vector-based alignment and scoring method described in Chapter 4.

Informed by the results in Chapter 2, Chapter 4 describes the development and benchmarking of a new, fully open-source implementation of HCIE, addressing key limitations of the previous version through a novel, vector-based alignment and scoring algorithm, and the use

of MoBiVic as a searchable library. The inclusion of user-defined exit-vectors, combined with a transparent and extensible Python codebase, improved both usability and reproducibility.

To validate the approach, a series of benchmarking studies were performed. A query structure corresponding to 3,5-disubstituted pyrazolopyridine (a heterocycle featuring in several recently approved kinase inhibitors) was searched against the MoBiVic database. Of the top 10 hits, two were known literature-reported bioisosteres, while six had never previously been described in this context, despite matching the original query in terms of hydrogen bonding patterns and exit-vector geometry. This indicated HCIE's capacity not only to return validated motifs but also to propose unexplored, synthetically accessible heterocycles as novel bioisosteres. A similar search using 2-pyridine, the most common aromatic nitrogen heterocycle in recent FDA-approved drugs, revealed that HCIE could enrich for known bioisosteres more effectively than an established Electroshape USR method, while also providing geometrical alignment information that facilitates downstream computational workflows such as virtual compound enumeration. As a consequence of these results, a series of 5,5-bicyclic aromatic heterocyclic bioisosteres of 2-pyridine that had never previously been characterised were proposed as an area of future work, and a synthetic route towards them outlined. A manuscript outlining the development of HCIE and these initial results has been submitted, and is under review at the time of writing.

To further explore the relevance of HCIE's scoring methodology, the tool was applied to five literature- and patent-derived series of NLRP3 inflammasome inhibitors. In each case, a reference ligand was aligned with its analogues, and total similarity scores (a weighted combination of electrostatic and shape similarity) were compared to experimental pIC_{50} values. Correlation coefficients ranged from 0.43 to 0.75, and the correlation was consistently improved by tuning the weightings of shape and electrostatics in the scoring function. These

findings support the hypothesis that bioactivity for this class of molecules is closely related to a combination of molecular shape and electrostatic potential, and show that HCIE can be adapted to prioritise similarity metrics relevant to a given target class or compound series.

Together, these benchmarking results support HCIE's dual capability to retrieve previously characterised bioisosteres and propose structurally novel yet functionally plausible replacements, thus expanding the pool of medicinally relevant aromatic heterocycles available for drug discovery.

The final two chapters described two streams of work that applied computational bioisosteric tools (including HCIE) in experimental settings, aiming to assess their real-world utility in early-stage medicinal chemistry projects. Chapter 5 involved exploratory synthetic attempts to validate novel ring-forming disconnections proposed by a domain-adapted molecular transformer model trained on heterocyclic cyclisations developed in the Duarte and Brennan Groups. Although the synthesis of new VEHICLe heterocycles was ultimately unsuccessful, these experiments provided key insight into the challenges faced when synthesising low molecular weight, polar ring systems. Difficulties included unfavourable polymerisation, failure of cyclisation due to the geometric constraints of intermediates such as hydrazones and oximes, and the analytical challenge in tracking highly polar small molecules. These findings highlighted the importance of geometric isomerism and molecular size and polarity in synthetic planning, and were used to inform the filtering rules in the MoBiVic library discussed in Chapter 3. Future work in this area should focus on more functionalised, derivatisable scaffolds, and strategies to promote isomerisation of reactive intermediates.

Chapter 6 described the use of HCIE to design and explore bioisosteric replacements in two active small molecule drug discovery projects in the CMD. The first involved replacing the

central heterocycle of a NUDT14 inhibitor identified based on a previous kinase inhibitor repurposing screen. Although the resulting compound maintained target selectivity, its potency was reduced, likely due to altered π-stacking interactions. The second project focused on the design of SUV4-20 inhibitors. This described ongoing synthesis, biological characterisation, and pharmacokinetic profiling of a set of compounds designed based on a virtual screening pipeline developed by researchers in the Brennan Group, ultimately leading to the discovery of a full inhibitor of SUV4-20 with a biochemically measured IC_{50} of 1.20 µM, comparable to that of the literature tool compound. A HCIE search identified several promising novel heterocyclic scaffolds, including one with a methoxy substituent as a potential bioisostere of the previously used dichlorophthalazine core with a lower lipophilicity. Ongoing docking studies, synthesis, and further biochemical assays will assess their viability as leads. Together, these projects demonstrate how HCIE can propose both novel and synthetically accessible bioisosteres that can be integrated into discovery pipelines.

The work presented in this thesis represents an exploration of aromatic heterocycles as bioisosteres in small molecule drug discovery, from their fundamental properties and medicinal utility to the development and real-world application of computational tools for their identification and prioritisation. The design, implementation, and evaluation of HCIE has demonstrated that electrostatic and shape-based similarity can serve as a powerful predictor of bioisosteric potential. HCIE provides a systematic, extensible software tool to uncover both known and previously unexplored aromatic heterocyclic scaffolds, thereby expanding the medicinal chemist's toolkit with new opportunities for potency, selectivity, and pharmacokinetic optimisation. By bridging cheminformatics and synthetic chemistry through benchmarking, database curation, and experimental validation, the work in this thesis has illustrated the central role of aromatic heterocycles in modern medicinal chemistry, but also

demonstrates how computational methods can guide and accelerate the search through novel chemical space. As the space of synthetically accessible molecules continues to grow alongside the need to improve the efficiency of drug discovery, tools like HCIE represent a step towards a systematic and data-driven approach to scaffold design, complementing the medicinal chemist's creativity and intuition, and presenting new opportunities for the discovery of better, safer therapeutics.

8 Experimental Details

8.1 Computational Details

Hardware and Computational Resources

Unless otherwise specified, all HCIE searches outlined in Chapters 2, 4, and 6, and all ETKDG optimisations and Gasteiger charge calculations were run on a MacBook Pro equipped with an Apple M1 Pro chip (8-core CPU, 16 GB of unified memory) running macOS 12.x. All xTB and DFT electronic structure calculations were executed on the Duarte Group Aleph cluster at the University of Oxford. The cluster includes the following nodes: eleven 24-core Intel Xeon Gold 6126 CPUs with 92.9 GB RAM, one 28-core Intel Xeon Gold 6132 CPU with 187.2 GB RAM, two 40-core Intel Xeon Gold 6230 CPUs with 187.2 GB RAM, one 40-core Intel Xeon Gold 5281R CPU with 92.9 GB RAM, and one 48-core Intel Xeon Gold 5318Y CPU with 93.5 GB RAM. Calculations were performed in batches, with each calculation assigned 8 cores unless otherwise stated. Multiple jobs were run concurrently depending on queue availability.

Software Details

All software packages and dependencies were managed using Conda (Anaconda), and environments were version-controlled to ensure reproducibility. The software versions and Python package versions used to derive the results outlined in Chapters 2, 4, and 6 are listed in Table 8.1.

Table 8.1. Software versions and Python packages used in Chapter 2 and in Chapters 4 and 6.

Software / Package	Chapter 2	Chapters 4 and 6
Python	3.9.6	3.12.7
xTB	6.4.0	6.6.1
ORCA	4.2.1	4.2.1
ShaEP	1.3.1	–
autodE	1.3.0	–
NumPy	1.21.1	2.1.2
Pandas	1.3.0	–
RDKit	2021.03.4	2024.09.01
SciPy	–	1.14.1
ODDT	0.7	–

8.2 General Experimental Details

Solvents and Reagents

All reactions were performed using flame-dried reaction vessels under an atmosphere of dried nitrogen unless stated otherwise. All solvents were purchased from commercial suppliers and used without further purification (HPLC or analytical grade). Anhydrous solvents were purchased from Acros Organics, and stored under a nitrogen atmosphere over activated molecular sieves. Standard vacuum line techniques were used throughout. Deionised water was provided by an Elgin DV 25 system. Organic solvents were dried during workup using anhydrous Na_2SO_4, or dried over a hydrophobic phase separator in the case of DCM. All other chemical reagents used were commercially available from Alfa Aesar, Fluorochem, and Sigma-Aldrich, and were used as supplied.

Purification and Chromatography

Thin layer chromatography (TLC) was carried out using aluminium-backed plates coated with 60 F_{254} silica gel sourced from Merck. Plates were visualised using UV light (short wavelength: 254 nm, long wavelength: 365 nm) or staining with ninhydrin (1 M in EtOH) or 1% aqueous $KMnO_4$ solution. Normal-phase silica gel column chromatography was carried out using Biotage Isolera One flash column chromatography systems. Compound analytical HPLC and low resolution mass spectrometry were carried out on a Waters LCMS system (Waters 2767 sample manager, Waters SFO system fluidics organiser, Waters 2545 binary gradient module, 2 × Waters 515 HPLC pumps, Waters 2998 photodiode array detector, Waters 2424 ELS detector, Waters SQ detector 2). Analytical separation used a Kinetex 5 µM EVO C18 column (100 mm × 3.0 mm, 100 Å) with a flow rate of 2 mL min^{-1} along a 3 minute gradient elution.The mobile phase consisted of **solvent A** (93% H_2O, 5% MeCN,

and 2% 0.5 M ammonium acetate adjusted to pH 6 with glacial acetic acid) and **solvent B** (18% H_2O, 80% MeCN, and 2% 0.5 M ammonium acetate adjusted to pH 6 with glacial acetic acid. The elution gradient was 95:5 (**A**:**B**) 0.35 min, then 95:5 (**A**:**B**) to 5:95 (**A**:**B**) over 1 min, then 5:95 (**A**:**B**) over 0.75 min, and then reversion back to 95:5 (**A**:**B**) over 0.1 min before running at 95:5 (**A**:**B**) for 0.8 min.

Mass Spectrometry

Low resolution mass spectra (LR-ESI-MS) were recorded on a Waters SQ Detector 2 (LCMS), and the data processed using Waters FractionLynx software. High resolution mass spectra (HR-ESI-MS) were recorded on a RapidFire 360 High Throughput MS system and processing was performed using Agilent MassHunter Workstation software.

NMR Spectroscopy

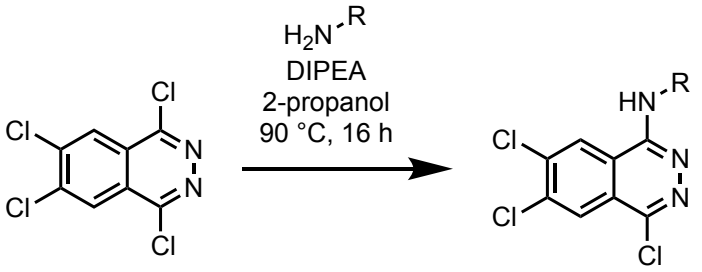
NMR spectra were recorded using a Bruker Avance 400 MHz spectrometer, with samples dissolved in the stated deuterated solvent. ^{13}C NMR spectroscopy was performed with broadband decoupling. Chemical shifts (δ) are quoted in parts per million (ppm) to the nearest 0.01 ppm for ^{1}H NMR spectra and to the nearest 0.1 ppm for ^{13}C NMR spectra (except where further precision is required for disambiguation) and referenced to the residual solvent peak. The data are presented as chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, quin = quintet, sext = sextet, dd = doublet of doublets, dt = doublet of triplets, m = multiplet, br. = broad resonance peak, app. = apparent, and any derivatives thereof), signal area integration, coupling constant (recorded in Hz and rounded to the nearest 0.1 Hz), and atom assignment. Two-dimensional NMR experiments (COSY, HSQC, HMBC, TOCSY) were used to aid in the assignment of ^{1}H and ^{13}C NMR spectra. All NMR spectra were processed using MestReNova v14.x software.

Reporting of Compounds

Systematic compound names were generated using ChemDraw v19.0.1.32 by PerkinElmer according to the guidelines specified by the International Union of Pure and Applied Chemistry (IUPAC). The atom numbering in displayed structures is for the purposes of spectral assignment and may not correspond to IUPAC numbering conventions. NMR spectral assignments follow the numbering schemes on each chemical structure.

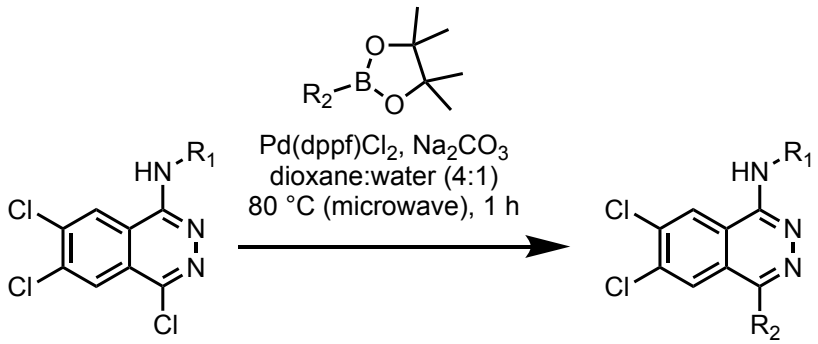
8.3 General Experimental Procedures

General Procedure A: Nucleophilic Aromatic Substitution



To a solution of 1,4,6,7-tetrachlorophthalazine (1 equiv.) in 2-propanol (0.1 M) was added amine (as a hydrochloride salt, 1 equiv.) and dry *N,N*-diisopropylethylamine (3 equiv.). The reaction was allowed to stir at 90 °C for 16 hours, then was diluted with 2-propanol (20 mL). The crude mixture was purified by flash column chromatography.

General Procedure B: Suzuki Coupling

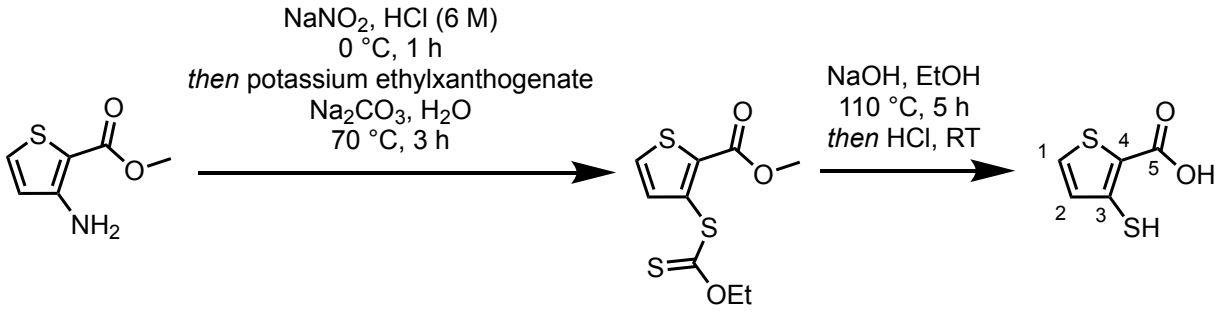


An oven-dried microwave vial equipped with stirrer bar was charged with 4,6,7-trichlorophthalazin-1-amine (1 equiv.), boronic acid pinacol ester (1.1 equiv.), sodium carbonate (3 equiv.), and [1,1-bis(diphenylphosphino) ferrocene] dichloropalladium(II) (0.2 equiv.). The vial was subsequently sealed, and evacuated & backfilled with dry nitrogen thrice. The solids were taken up in dry 1,4-dioxane (10 mL mmol^{-1}) and degassed water (2.5 mL mmol^{-1}), and the vial again evacuated & backfilled with nitrogen thrice. The resulting suspension was

heated under microwave irradiation for 1 hour at 80 °C, then cooled to room temperature and diluted with 1,4-dioxane (50 mL mmol^{-1}). The resulting suspension was passed through a plug of celite, and the solvent removed *in vacuo*. The crude mixture was purified by flash column chromatography.

8.4 Compound Synthesis and Characterisation

3-Mercaptothiophene-2-carboxylic acid, 5.1



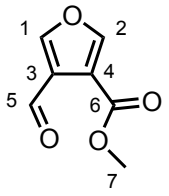
From a procedure documented by Sheriff *et al.*[401] 3-amino-2-methoxycarbonylthiophene (523 mg, 3.33 mmol) was added portionwise to hydrochloric acid (6.0 M in H_2O, 4.0 mL, 24 mmol), forming a pale-cream suspension, to which was added dropwise with stirring on ice a solution of sodium nitrite (270 mg, 3.91 mmol) in water (3 mL). The resulting solution was stirred on ice for one hour, during which it turned yellow, before being poured carefully into a solution of potassium ethylxanthogenate (594 mg, 3.71 mmol) and sodium carbonate (3.22 g, 30.4 mmol) in water (25 mL) after which it took a brilliant-yellow colour.[a] This solution was stirred at 60 °C for 40 minutes, before being heated under reflux at 70 °C for 2 hours, during which a deep-orange solution with a putrid odour formed. This was cooled to room temperature, diluted with water (15 mL), and extracted with ethyl acetate (3 × 20 mL). The combined organic layers were washed with water (15 mL), brine (10 mL), dried over sodium sulfate, and the solvent removed *in vacuo* to give the **intermediate** as a brown amorphous solid with a pungent odour, which was carried forward immediately without further purification (730 mg, 2.79 mmol, 84% yield).

[a]Great care must be taken with the addition to avoid significant foaming.

The xanthogenate **intermediate** (652 mg, 2.49 mmol) was dissolved in ethanol (6 mL), and sodium hydroxide solution (5.0 M in water, 22 mL, 100 mmol) was added with stirring. The resulting coffee-coloured solution was heated under reflux at 110 °C for two hours. After cooling, the volatiles were removed under reduced pressure, and the resulting pale-orange solution was extracted with ether (2 × 15 mL) before being acidified to pH 3 with hydrochloric acid (6 M solution in H_2O), at which point a yellow precipitate formed. This suspension was extracted with ethyl acetate (3 × 20 mL), and the combined organic layers washed with brine (15 mL), dried over sodium sulfate, and the solvent removed under reduced pressure to give the **title compound** as a brown amorphous solid with a foul odour (420 mg, 2.62 mmol, overall 84% yield).

^{1}**H NMR** (400 MHz, DMSO−d_6) δ 7.84 (d, 1H, J = 5.2 Hz, C_1–**H**), 7.17 (d, 1H, J = 5.2 Hz, C_2–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 163.0 (C_5), 141.9 (C_4), 132.7 (C_1), 127.5 (C_2), 124.6 (C_3); **LRMS** (*m/z* -ESI): found $[M]^-$: 160.2, calculated for $C_5H_4O_2S_2$ $[M]^-$: 160.0; All data in accordance with literature.[401]

Methyl 4-formylfuran-3-carboxylate, 5.9

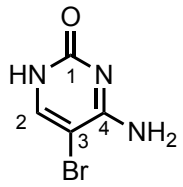


From a procedure documented by Davies *et al.*[402] A solution of oxalyl chloride (0.18 mL, 2.1 mmol) in dry dichloromethane (8 mL) was cooled to -78 °C and DMSO (0.29 mL, 4.1 mmol) was added dropwise with stirring. When the addition was complete the resulting solution was stirred at -78 °C for 30 minutes, before a solution of methyl 4-(hydroxymethyl)furan-3-

carboxylate **5.22** (280 mg, 1.79 mmol) in dry dichloromethane (10 mL) was added dropwise with stirring. The solution was stirred for a further 30 minutes at -78 °C before triethylamine (1.20 mL, 8.97 mmol) was added dropwise with stirring, and was allowed to warm to room temperature once addition was complete. The reaction was stirred at room temperature overnight, and then diluted with water (10 mL) and partitioned. The aqueous layer was washed with dichloromethane (3 × 15 mL) and the combined organic layers washed with water (10 mL), brine (10 mL), dried over a phase separator, and the solvent removed *in vacuo* to afford the **title compound** as a brown oil (97 mg, 0.63 mmol, 35% yield).

$\mathbf{R}_f$ (20 % EtOAc in cyclohexane) 0.42; ^{1}**H NMR** (400 MHz, $CDCl_3$) δ 10.35 (s, 1H, C_5–**H**), 8.05 (d, 1H, J = 1.6 Hz, $C_{1/2}$–**H**), 8.02 (d, 1H, J = 1.6 Hz, $C_{1/2}$–**H**), 3.90 (s, 3H, C_7–**H**); ^{13}**C NMR** (100 MHz, $CDCl_3$) δ 186.7 (C_5), 162.7 (C_6), 149.4 ($C_{1/2}$), 148.0 ($C_{1/2}$), 125.6 ($C_{3/4}$), 117.6 ($C_{3/4}$), 52.1 (C_7); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 155.3, calculated for $C_7H_5O_4$ $[M+H]^+$: 155.0; All data in accordance with literature.[402]

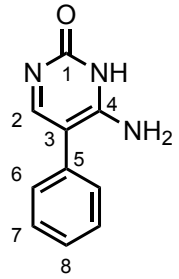
4-Amino-5-bromopyrimidin-2(1H)-one, 5.16



To a stirred suspension of **cytosine** (10.18 g, 91.63 mmol) in *N,N*-dimethylformamide (100 mL) was added *N*-bromosuccinimide (16.32 g, 91.70 mmol), and the resulting bright yellow suspension stirred at room temperature overnight. The reaction was filtered under reduced pressure, and the filtrand washed with water (3 × 20 mL) and dried *in vacuo* overnight to give the **title compound** as an off-white amorphous solid (12.03 g, 63.32 mmol, 69% yield).

^{1}H NMR (400 MHz, DMSO−d_6) δ 10.83 (br. s, 1H, CON**H**), 7.74 (s, 1H, C_2–**H**), 6.83 (br. s, 2H, N**H**); **^{13}C NMR** (100 MHz, DMSO−d_6) δ 163.2 (C_1), 156.0 (C_4), 144.5 (C_3), 85.7 (C_2); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 192.0, calculated for $C_4H_4{}^{81}BrN_3O$ $[M+H]^+$: 192.0; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 189.9594, calculated for $C_4H_5{}^{81}BrN_3O^+$: 189.9611.

6-Amino-5-phenylpyrimidin-2(1H)-one, 5.17

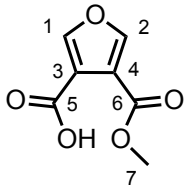


5.16 (132 mg, 69.5 µmol), phenylboronic acid pinacol ester (287 mg, 1.41 mmol), and tribasic potassium phosphate (461 mg, 2.18 mmol) were added to a microwave vial, which was subsequently evacuated and backfilled with dry nitrogen thrice. 1,4-dioxane (3 mL) and water (3 mL) were added, and the resulting suspension again evacuated and backfilled with dry nitrogen thrice before XPhos Pd G2 (54 mg, 73 µmol) was added, and the resulting suspension was heated for one hour at 100 °C in a microwave. The resulting brown suspension was diluted with 1,4-dioxane (10 mL) and filtered through celite. Purification by flash column chromatography (SiO_2, 10% to 20% methanol in DCM) provided the **title compound** as a brown amorphous solid (26 mg, 0.14 mmol, 20% yield).

^{1}H NMR (400 MHz, DMSO−d_6) δ 8.02 (s, 1H, C_2–**H**), 7.82–7.75 (m, 1H, C_8–**H**), 7.46–7.35 (m, 2H, $C_{6/7}$–**H**), 7.35–7.26 (m, 4H, $C_{6/7}$–**H**, N**H**); **^{13}C NMR** (100 MHz, DMSO−d_6) δ 164.2 (C_1), 156.1 (C_4), 134.0 (C_8), 130.0 (C_5), 128.9 ($C_{6/7}$), 128.8 ($C_{6/7}$), 127.4 C_2); **LRMS**

(*m/z* +ESI): found $[M+H]^+$: 188.4, calculated for $C_{10}H_9N_3O$ $[M+H]^+$: 188.1; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 188.0839, calculated for $C_{10}H_{10}N_3O^+$: 188.0818.

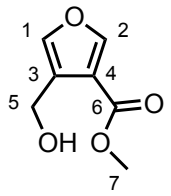
4-(Methoxycarbonyl)furan-3-carboxylic acid, 5.20



Adapted from a route by Mercogliano *et al.*[356] To a solution of dimethylfuran-3,4-dicarboxylate (2.11 g, 11.4 mmol) in methanol (60 mL) was added potassium hydroxide solution (2.0 M in H_2O, 8.5 mL, 17 mmol) and the resulting solution stirred at room temperature for 1.5 hours. The solution was then concentrated *in vacuo* and the residue taken up in water (50 mL) and washed with diethyl ether (10 mL). The aqueous phase was then acidified to pH 1 by the addition of HCl solution (1 M in H_2O), which caused the formation of a white precipitate. This was isolated by filtration, and washed with ice-cold water (10 mL) and diethyl ether (10 mL), before being dried *in vacuo* overnight, providing the **title compound** as a white amorphous solid (1.47 g, 8.62 mmol, 75% yield).

^{1}**H NMR** (400 MHz, CD_3OD) δ 8.26 (d, 1H, J = 1.7 Hz, $C_{1/2}$–**H**), 8.24 (d, 1H, J = 1.7 Hz, $C_{1/2}$–**H**), 3.91 (s, 3H, C_7–**H**); ^{13}**C NMR** (100 MHz, CD_3OD) δ 166.1 (C_5), 164.3 (C_6), 152.2 ($C_{1/2}$), 151.6 ($C_{1/2}$), 119.8 ($C_{3/4}$), 118.1 ($C_{3/4}$), 53.2 (C_7); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 171.3, calculated for $C_7H_6O_5$ $[M+H]^+$: 171.0; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 171.0291, calculated for $C_7H_7O_5^+$: 171.0288. All data in accordance with literature.[356]

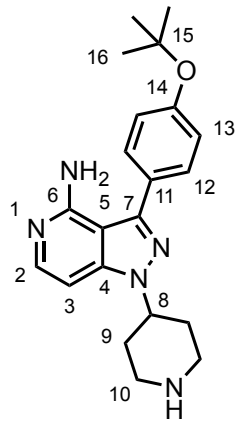
Methyl 4-(hydroxymethyl)furan-3-carboxylate, 5.22



Adapted from a procedure described by Hawker and Silverman.[403] To a stirred solution of 4-(methoxycarbonyl)furan-3-carboxylic acid **5.20** (1.03 g, 6.06 mmol) in dry THF (24 mL) cooled on ice was added borane in complex with dimethylsulfide (0.86 mL, 9.1 mmol) dropwise with stirring. The reaction mixture was allowed to warm to room temperature and stirred overnight under an atmosphere of dry nitrogen, during which a white precipitate formed. The resulting suspension was cooled on ice and water (5 mL) was added dropwise with stirring until effervescence was observed to cease. The reaction mixture was concentrated *in vacuo*, and the residue partitioned between a saturated aqueous solution of $NaHCO_3$ (15 mL) and diethyl ether (15 mL). The aqueous layer was further washed with diethyl ether (2 × 20 mL), and the combined ethereal layers washed with brine (15 mL) and dried over sodium sulfate. Evaporation of the solvent under reduced pressure afforded the **title compound** as a clear oil, which was taken forward without further purification (506 mg, 3.24 mmol, 54% yield).

$\mathbf{R}_f$ (40 % EtOAc in cyclohexane) 0.51; ^{1}**H NMR** (400 MHz, $CDCl_3$) δ 7.98 (s, 1H, C_2–**H**), 7.40 (s, 1H, C_1–**H**), 4.62 (s, 2H, C_5–**H**), 3.86 (s, 3H, C_7–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 163.2 (C_6), 149.5 (C_2), 141.9 (C_1), 125.5 (C_4), 116.78 (C_3), 55.0 (C_5), 51.5 (C_7); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 157.2, calculated for $C_7H_8O_4$ $[M+H]^+$: 157.0; All data in accordance with literature.[403]

3-(4-(Tert-butoxy)phenyl)-1-(piperidin-4-yl)-1H-pyrazolo[4,3-c]pyridin-4-amine, 6.3

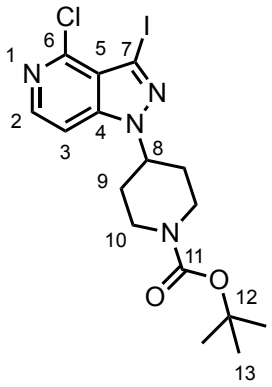


To a stirred, degassed solution of 3-iodo-1-(piperidin-4-yl)-1*H*-indazol-4-amine **6.9** (137 mg, 40.0 µmol), 2-(4-*tert*-butoxy)phenyl boronic acid pinacol ester (289 mg, 1.05 mmol), and potassium carbonate (348 mg, 2.52 mmol) in dioxane (4 mL) and water (1 mL) was added [1,1'-Bis(diphenylphosphino)ferrocene]dichloropalladium (II) complex with dichloromethane (67 mg, 8.1 µmol) under an atmosphere of nitrogen. The reaction was heated under microwave irradiation at 100 °C for 35 minutes, after which the reaction was cooled to room temperature and filtered through a pad of celite. The solvent was removed *in vacuo* and the residue taken up in water (10 mL), extracted with ethyl acetate (3 × 10 mL), and the combined organic layers washed with brine (10 mL) and dried over sodium sulfate. Purification of the crude product by preparative HPLC provided the **title compound** as a brown oil (2.4 mg, 6.6 µmol, 7% yield).

^{1}**H NMR** (400 MHz, $CDCl_3$) δ 7.62 (d, 1H, J = 6.8 Hz, C_2–**H**), 7.53 (app. dt, 2H, J = 8.6, 2.1 Hz, C_{12}–**H**), 7.14 (dt, 2H, J = 8.6, 2.1 Hz, C_{13}–**H**), 6.75 (d, 1H, J = 6.9 Hz, C_3–**H**), 4.58-4.44 (m, 1H, C_8–**H**), 3.45 (app. dt, 2H, J = 13.1, 3.3 Hz, C_{10}–**H**), 3.03-2.88 (m, 2H, C_{10}'–**H**), 2.45-2.30 (m, 2H, C_9–**H**), 2.19-2.11 (m, 2H, C_9'–**H**), 1.41 (s, 9H, C_{16}–**H**); ^{13}**C NMR** (100

MHz, MeOD) δ 156.3 (C_{14}), 153.7 (C_6), 145.1 (C_{11}), 144.6 (C_7), 143.2 (C_4), 130.0 (C_{12}), 128.3 (C), 124.6 (C_{13}), 106.1 (C_5), 96.4 (C_3), 79.2 (C_{15}), 57.3 (C_8), 46.2 (C_{10}), 33.1 (C_9), 29.0 (C_{16}); **LRMS** (*m/z* +ESI): found $[M+H]^+$, 366.6, calculated for $C_{21}H_{28}N_5O^+$: 366.2; **HRMS** (*m/z*): found M^+: 366.2283. Calculated for $C_{21}H_{27}N_5O^+$: 366.2288.

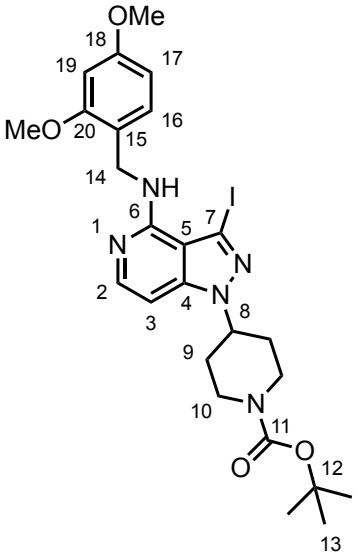
Tert-butyl 4-(4-chloro-3-iodo-1H-pyrazolo[4,3-c]pyridin-1-yl)piperidine-1-carboxylate, 6.7



To a suspension of polymer-supported triphenylphosphine (418 mg, 1.89 mmol g^{-1} loading, 790 µmol) in dry THF (4.0 mL) on ice was added diisopropyl azodicarboxylate (156 µL, 794 µmol) dropwise with stirring, and the resulting suspension stirred on ice for 15 minutes. 1-*tert*-butoxycarbonyl-4-hydroxypiperidine (114 mg, 566 µmol) was added, and the suspension stirred on ice for a further 5 minutes before addition of 4-chloro-3-iodo-1*H*-pyrazolo[4,3*c*]pyridine (101 mg, 361 µmol), and the resulting suspension stirred at room temperature overnight. The solution was then filtered under reduced pressure, and the solvent removed *in vacuo* to give a yellow oil which was purified by flash column chromatography (SiO_2, eluent: 20% EtOAc in cyclohexane) providing the **title compound** as an off-white amorphous solid (84 mg, 0.18 mmol, 51%).

$\mathbf{R}_f$ (30% EtOAc in cyclohexane) 0.37; ^{1}**H NMR** (400 MHz, $CDCl_3$) δ 7.99 (d, 1H, J = 6.2 Hz, C_2–**H**), 7.49 (d, 1H, J = 6.2 Hz, C_3–**H**), 4.85 (tt, 1H, J = 11.4, 4.1 Hz, C_8–**H**), 4.33 (br. s, 2H, C_{10}–**H**), 2.94 (br. s, 2H, C_{10}'–**H**), 2.32 - 2.17 (m, 2H, C_9–**H**), 2.04-1.96 (m, 2H, C_9'–**H**), 1.47 (s, 9H, C_{13}–**H**); ^{13}**C NMR** (100 MHz, $CDCl_3$) δ 154.5 (C_{11}), 151.7 (C_6), 145.2 (C_4), 141.2 (C_2), 120.8 (C_5), 112.0 (C_3), 80.3 (C_{12}), 77.7 (C_7), 61.6 (C_8), 42.8 (br., C_{10}) 31.9 (C_9), 28.5 (C_{13}); **LRMS** (*m/z* +ESI): found $[M+H]^+$, 463.1. Calculated for $C_{16}H_{20}^{35}ClIN_4O_2$ $[M+H]^+$: 463.0. **HRMS** (*m/z* +ESI): found $[M+H]^+$: 463.0419, calculated for $C_{16}H_{21}^{35}ClIN_4O_2^+$: 463.0392.

Tert-butyl 4-(4-((2,4-dimethoxybenzyl)amino)-3-iodo-1H-pyrazolo[4,3-c]pyridin-1-yl)piperidine-1-carboxylate, 6.8



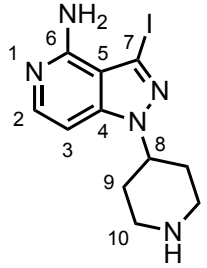
To a stirred solution of 4-chloro-3-iodo-1-(piperidin-4-yl)-1*H*-pyrazolo[4,3-*c*]pyridine (275 mg, 594 µmol) in dry dioxane (5 mL) was added 1,3-dimethoxybenzylamine (0.45 mL, 3.0 mmol) and triethylamine (1.00 mL, 7.17 mmol), and the resulting solution heated under microwave irradiation at 180 °C for one hour. The solvent was then removed *in vacuo* and the

resulting residue taken up in water (10 mL) and extracted with ethyl acetate (3 × 10 mL). The combined organic layers were washed with water (10 mL), brine (10 mL), dried over sodium sulfate, and concentrated under reduced pressure. Purification of the crude mixture by flash column chromatography (SiO_2, 20% to 60% EtOAc in cyclohexane) provided the **title compound** as a white amorphous solid, which was carried forward without further purification (265 mg, 0.447 mmol, 75% yield).

R_f (60% EtOAc in cyclohexane) 0.40; **^{1}H NMR** (400 MHz, MeOD) δ 7.74 (d, 1H, J = 6.3 Hz, C_2–**H**), 7.23 (d, 1H, J = 8.3 Hz, C_{16}–**H**), 6.83 (d, 1H, J = 6.3 Hz, C_3–**H**), 6.57 (d, J = 2.5 Hz, C_{19}–**H**), 6.45 (dd, 1H, J = 8.3, 2.5 Hz, C_{17}–**H**), 4.61 (s, 2H, C_{14}–**H**), 4.64-4.54 (m, 1H, C_8–**H**), 4.20 (d, 2H, J = 13.6 Hz, C_{10}–**H**), 3.90 (s, 3H, O**Me**), 3.77 (s, 3H, O**Me**), 2.98 (br s, 2H, C_{10}'–**H**), 2.02 (app. qd, 2H, J = 12.4, 4.4 Hz, C_9–**H**), 1.95-1.84 (m, 2H, C_9'–**H**), 1.48 (s, 9H, C_{13}–**H**); **^{13}C NMR** (100 MHz, MeOD) δ 162.3 (C_{18}), 160.2 (C_{20}), 156.4 (C_{11}), 153.9 (C_6), 145.6 ($C_{4/5}$), 144.3 (C_2), 131.2 (C_{16}), 120.0 (C_{15}), 111.6 ($C_{4/5}$), 105.2 (C_{17}), 99.4 (C_{19}), 96.2 (C_3), 88.7 (C_7), 81.3 (C_{12}), 57.4 (C_8), 56.1 (O**Me**), 55.8 (O**Me**), 41.7 (C_{14}), 32.5 (C_9), 28.7 (C_{13}); **LRMS** (*m/z* +ESI): found $[M+H]^+$, 594.2, calculated for

$C_{25}H_{32}IN_5O_4$ $[M+H]_+$: 594.2; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 594.1657, calculated for $C_{25}H_{33}IN_5O_4^+$: 594.1572.

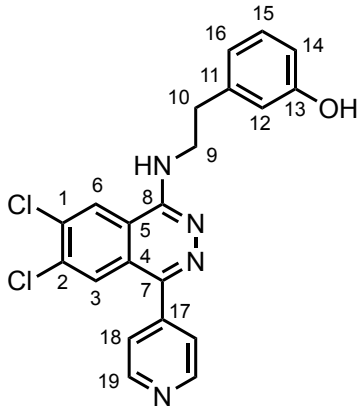
3-Iodo-1-(piperidin-4-yl)-1H-indazol-4-amine, 6.9



To a stirred solution of *tert*-butyl 4-(4-((2,4-dimethoxybenzyl)amino)-3-iodo-1*H*-pyrazolo[4,3-*c*]pyridin-1-yl)piperidine-1-carboxylate **6.8** (233 mg, 393 µmol) in DCM (12 mL) was added anisole (0.43 mL, 3.9 mmol) and trifluoroacetic acid (0.65 mL, 8.5 mmol), at which point the solution turned bright yellow. This was stirred at room temperature overnight, after which it turned a pleasing pink colour. The solution was concentrated *in vacuo* and the residue taken up in methanol (10 mL) and purified by strong cation exchange chromatography, eluting with 2 M methanolic ammonia. The volatiles were removed *in vacuo* providing the **title compound** as a yellow amorphous solid (118 mg, 344 µmol, 88% yield).

^{1}**H NMR** (400 MHz, MeOD) δ 7.16 (d, 1H, J = 6.4 Hz, C_2–**H**), 6.38 (d, 1H, J = 6.38 Hz, C_3–**H**), 4.06-3.94 (m, 1H, C_8–**H**), 2.70 (app. dt, 2H, J = 12.8, 3.6 Hz, C_{10}–**H**), 2.28 (app. td, J = 12.8, 2.7 Hz, C_{10}'–**H**), 1.62 (app. qd, 2H, J = 12.8, 4.2 Hz, C_9–**H**), 1.45 (app. dt, 2H, J = 12.4, 4.2, 2.7 Hz, C_9'–**H**); ^{13}**C NMR** (100 MHz, MeOD) δ 152.9 (C_6), 143.9 ($C_{4/5}$), 142.0 (C_2), 110.0 ($C_{4/5}$), 95.3 (C_3), 87.3 (C_7), 56.1 (C_{12}), 44.3 (C_{10}), 31.5 (C_9); **LRMS** (*m/z* +ESI): found $[M+H]^+$, 344.1, calculated for $C_{11}H_{15}IN_5^+$: 344.0; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 344.0368, calculated for $C_{11}H_{15}IN_5^+$: 344.0367.

3-(2-((6,7-Dichloro-4-(pyridin-4-yl)phthalazin-1-yl)amino)ethyl)phenol, 6.16

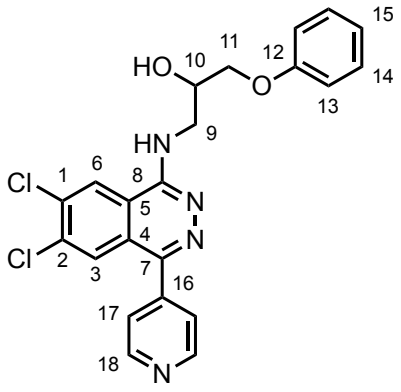


An oven-dried microwave vial equipped with a stirrer bead was charged with 3-(2-((4,6,7-trichlorophthalazin-1-yl)amino)ethyl)phenol **6.30** (35 mg, 94 µmol), 4-pyridineboronic acid pinacol ester (30 mg, 0.15 mmol), potassium carbonate (32 mg, 0.24 mmol), and polymer-supported tetrakis(triphenylphosphine)palladium(0) (47 mg, 0.40 mmol g^{-1} loading, 14 µmol). The vial was sealed, and evacuated & backfilled with dry nitrogen thrice. The solids were then taken up in dry THF (0.5 mL) and water (0.3 mL), and the vial again evacuated & backfilled with dry nitrogen thrice. The resulting suspension was heated at 80 °C for 48 hours under an atmosphere of nitrogen, before being cooled to room temperature and the volatiles removed *in vacuo*. The resulting residue was taken up in a solution of 25% dichloromethane in methanol (10 mL) and filtered to give a red solution which was purified by preparative HPLC to provide the **title compound** as an off-white amorphous solid (3.0 mg, 7.0 µmol, 8% yield).

^{1}**H NMR** (400 MHz, DMSO−d_6) δ 9.31 (br. s, 1H, O**H**), 8.78-8.75 (m, 3H, C_6–**H** & C_{19}–**H**), 8.05 (t, 1H, J = 5.5 Hz, N**H**), 7.96 (s, 1H, C_3–**H**), 7.72-7.66 (m, 2H, C_{18}–**H**), 7.10 (t, 1H, J = 7.0 Hz, C_{15}–**H**), 6.74-6.68 (m, 2H, C_{12}–**H** & C_{16}–**H**), 6.62-6.58 (m, 1H, C_{14}–**H**),

3.80 (app. q, 2H, J = 7.0 Hz, C_9–**H**), 2.94 (t, 2H, J = 7.5 Hz, C_{10}–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 157.4 (C_{13}), 152.4 (C_8), 150.0 (C_{19}), 147.5 (C_7), 143.8 (C_{17}), 141.1 (C_{11}), 135.0 (C_1), 134.2 (C_2), 129.3 (C_{15}), 126.6 (C_3), 125.0 (C_6), 124.6 ($C_{4/5}$), 124.2 (C_{18}), 119.3 ($C_{16/12}$), 117.1 ($C_{4/5}$), 115.6 ($C_{12/16}$), 113.1 (C_{14}), 42.8 (C_9), 34.4 (C_8); **LRMS** (m/z +ESI): found $[M+H]^+$: 411.2, calculated for $C_{21}H_{16}^{35}Cl_2N_4O$ $[M+H]^+$: 411.1; **HRMS** (m/z +ESI): found $[M+H]^+$: 411.0792. Calculated for $C_{21}H_{17}^{35}Cl_2N_4O^+$: 411.0774.

1-((6,7-Dichloro-4-(pyridin-4-yl)phthalazin-1-yl)amino)-3-phenoxypropan-2-ol, 6.17

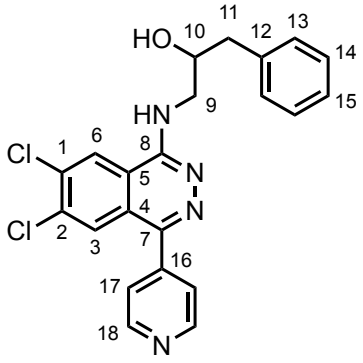


An oven-dried microwave vial equipped with a stirrer bead was charged with 1-phenoxy-3-((4,6,7-trichlorophthalazin-1-yl)amino)propan-2-ol **6.31** (39 mg, 0.10 mmol), 4-pyridineboronic acid pinacol ester (32 mg, 0.16 mmol), potassium carbonate (36 mg, 0.26 mmol), and tetrakis(triphenylphosphine)palladium(0) (27 mg, 23 µmol). The vial was sealed, and evacuated & backfilled with dry nitrogen thrice. The solids were then taken up in dry THF (1 mL) and water (0.2 mL), and the vial again evacuated & backfilled with dry nitrogen thrice. The resulting suspension was heated at 80 °C for 16 hours under an atmosphere of nitrogen, before being cooled to room temperature and the volatiles removed *in vacuo*. The

resulting residue was taken up in a saturated ethanolic solution (10% methanol in ethanol) of zinc(II) chloride and shaken at room temperature overnight. The resulting suspension was filtered under reduced pressure, and the yellow filtrate concentrated *in vacuo* and taken up in methanol (10 mL), and purified by strong cation exchange chromatography, eluting with 2 M methanolic ammonia. The volatiles were again removed *in vacuo*, and the crude product further purified by preparative HPLC to afford the **title compound** as a pale-yellow amorphous solid (7.0 mg, 16 µmol, 16% yield).

^{1}**H NMR** (400 MHz, DMSO−d_6) δ 8.01 (s, 1H, C_6–**H**), 7.97-7.89 (m, 2H, C_{18}–H), 7.22 (br. s, 1H, O**H**), 7.10 (s, 1H, C_3–**H**), 6.86-6.80 (m, 2H, C_{17}–**H**), 6.48-6.39 (m, 2H, C_{14}–**H**), 6.15-6.04 (m, 3H, C_{13}–**H** & C_{15}–**H**), 4.65 (br. s, 1H, N**H**), 3.45 (app. quin, 1H, J = 4.9 Hz, C_{10}–**H**), 3.22 (dd, 1H, J = 10.0, 4.3 Hz, C_{11}–**H**), 3.16 (dd, 1H J = 10.0, 5.8 Hz, C_{11}'–**H**), 3.01 (app. dt, 1H, J = 10.9, 5 Hz, C_9–**H**), 2.85 (app. dt, 1H, J = 10.9, 5.6 Hz, C_9'–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 159.1 (C_{12}), 153.1 (C_8), 150.5 (C_{18}), 148.1 (C_7), 144.2 (C_{16}), 135.5 (C_1), 134.6 (C_2), 129.9 (C_{14}), 127.0 (C_3), 125.7 (C_6), 125.1 ($C_{4/5}$), 124.6 (C_{17}), 121.0 (C_{15}), 117.7 ($C_{5/4}$), 71.0 (C_{11}), 67.5 (C_{10}), 45.4 (C_9); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 441.1, calculated for $C_{21}H_{16}^{35}Cl_2N_4O$ $[M+H]^+$: 441.1; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 441.0872, calculated for $C_{21}H_{17}^{35}Cl_2N_4O^+$: 441.0880.

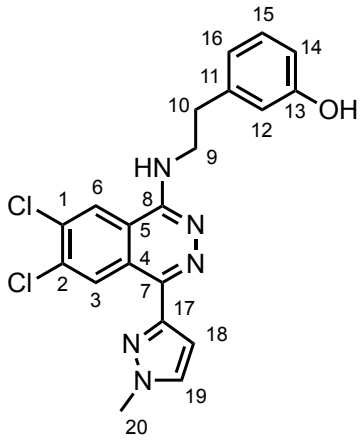
1-((6,7-Dichloro-4-(pyridin-4-yl)phthalazin-1-yl)amino)-3-phenylpropan-2-ol, 6.18



The **title compound** was prepared from 1-phenyl-3-((4,6,7-trichlorophthalazin-1-yl)amino)-propan-2-ol **6.32** (47 mg, 0.12 mmol) and 4-pyridineboronic acid pinacol ester (28 mg, 0.14 mmol) according to **General Procedure B**. Purification by preparative HPLC gave the **title compound** as a pale-yellow amorphous solid (4.8 mg, 11 µmol, 9% yield).

^{1}**H NMR** (400 MHz, DMSO−d_6) δ 8.86 (s, 1H, C_6–**H**), 8.79-8.74 (m, 2H, C_{18}–**H**), 8.07 (br. s, 1H, O**H**), 7.95 (s, 1H, C_3–**H**), 7.71-7.65 (m, 2H, C_{17}–**H**), 7.31-7.22 (m, 4H, C_{13}–**H** & C_{14}–**H**), 7.17 (tt, 1H, J = 5.3, 3.3 Hz, C_{15}–**H**), 5.14 (br. s, 1H, N**H**), 4.15 (app. tt, 1H, J = 7.5, 4.8 Hz, C_{10}–**H**), 3.71 (app. dt, 1H, J = 13.4, 5.1 Hz, C_9–**H**), 3.54 (ddd, 1H, J = 13.4, 7.5, 5.2 Hz, C_9'–**H**), 2.86 (dd, 1H, J = 13.7, 5.0 Hz, C_{11}–**H**), 2.74 (dd, 1H, J = 13.7, 7.6 Hz, C_{11}'–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 153.0 (C_8), 150.4 (C_{18}), 148.0 (C_7), 144.1 (C_{16}), 139.8 (C_{12}), 134.7 ($C_{4/5}$), 129.9 (C_{13}), 128.5 (C_{14}), 127.1 (C_3), 126.3 (C_{27}), 125.9 (C_6), 125.2 ($C_{5/4}$), 124.7 (C_{17}), 69.8 (C_{10}), 48.2 (C_9), 41.8 (C_{11}); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 425.3, calculated for $C_{22}H_{18}^{35}Cl_2N_4O$ $[M+H]^+$: 425.1; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 425.0930. Calculated for $C_{22}H_{19}^{35}Cl_2N_4O^+$: 425.0930.

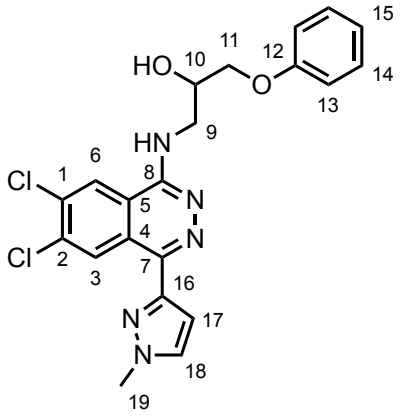
3-(2-((6,7-Dichloro-4-(1-methyl-1H-pyrazol-3-yl)phthalazin-1-yl)amino)ethyl)phenol, 6.21



The **title compound** was prepared from 3-(2-((4,6,7-trichlorophthalazin-1-yl)amino)ethyl)-phenol **6.30** (26 mg, 71 µmol) and 1-methyl-1*H*-pyrazole-3-boronic acid pinacol ester (17 mg, 81 µmol) according to **General Procedure B**. Purification of the crude product by preparative HPLC provided the **title compound** as a pale-red amorphous solid (6.9 mg, 17 µmol, 23% yield).

$\mathbf{R}_f$ (5% methanol in DCM) 0.12; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 9.50 (s, 1H, C_6–**H**), 8.69 (s, 1H, C_3–**H**), 7.87 (t, 1H, J = 5.5 Hz, N**H**), 7.84 (d, 1H, J = 2.2 Hz, C_{19}–**H**), 7.09 (t, 1H, J = 7.9 Hz, C_{15}–**H**), 6.90 (d, 1H, J = 2.2 Hz, C_{18}–**H**), 6.73-6.67 (m, 2H, C_{12}–**H** & C_{16}–**H**), 6.60 (ddd, 1H, J = 7.9, 2.3, 1.1 Hz, C_{14}–**H**), 4.00 (s, 3H, C_{20}–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 157.5 (C_{13}), 151.8 (C_8), 149.6 (C_{17}), 141.8 (C_7), 141.1 (C_{11}), 134.5 (C_1), 133.6 (C_2), 131.8 (C_{19}), 129.3 (C_{15}), 128.4 (C_3), 124.6 (C_6), 124.3, 119.2 (C_{16}), 117.4, 115.6 (C_{12}), 113.1 (C_{14}), 105.4 (C_{18}), 42.8 (C_9), 34.5 (C_{10}); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 414.2, calculated for $C_{20}H_{17}^{35}Cl_2N_5O$ $[M+H]^+$: 414.1; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 414.0898. Calculated for $C_{22}H_{19}^{35}Cl_2N_4O^+$: 414.0883.

1-((6,7-Dichloro-4-(1-methyl-1H-pyrazol-3-yl)phthalazin-1-yl)amino(-3-phenoxypropan-2-ol, 6.22



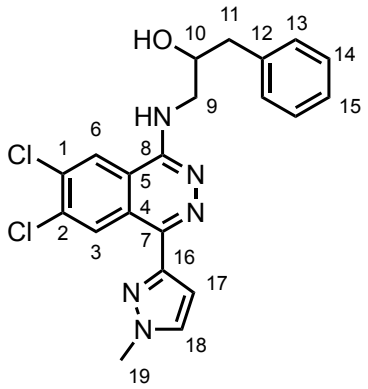
The **title compound** was prepared from 1-phenoxy-3-((4,6,7-trichlorophthalazin-1-yl)amino)-propan-2-ol **6.31** (51 mg, 0.13 mmol) and 1-methyl-1*H*-pyrazole-3-boronic acid pinacol ester (30 mg, 0.15 mmol) according to **General Procedure B**. Purification of the crude mixture by flash column chromatography (SiO_2, 50% to 100% EtOAc in cyclohexane) provided the *title compound* as a pale-red amorphous solid (33 mg, 74 µmol, 57% yield).

$\mathbf{R}_f$ (EtOAc) 0.47; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 9.49 (s, 1H, C_3–**H**), 8.75 (s, 1H, C_6–**H**), 7.86 (t, 1H, J = 5.5 Hz, N**H**), 7.83 (d, 1H, J = 2.3 Hz, C_{18}–**H**), 7.27 (td, 2H, J = 7.4, 1.3 Hz, C_{14}–**H**), 6.94 (td, 2H, J = 7.4, 1.1 Hz, C_{13}–**H**), 6.88 (d, 1H, J = 2.3 Hz, C_{17}–**H**), 5.45 (br. d, 1H, J = 4.5 Hz, O**H**), 4.30-4.25 (m, 1H, C_{10}–**H**), 4.06 (dd, 1H, J = 10.0, 4.3 Hz, C_{11}–**H**), 4.02 (dd, 1H, J = 10.0 Hz, C_{11}'–**H**)[b], 4.00 (s, 3H, C_{19}–**H**), 3.82 (app. dt, 1H, J = 13.4, 5.5 Hz, C_9–**H**), 3.67 (ddd, 1H, J = 13.4, 6.9, 5.5 Hz, C_9'–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 159.2 (C_{12}), 152.7 (C_8), 150.0 (C_{16}), 142.4 (C_7), 135.0 (C_1), 134.1 (C_2), 132.3 (C_{18}), 129.9 (C_{14}), 128.8 (C_3), 125.3 (C_6), 124.8 ($C_{4/5}$), 121.0 (C_{15}), 118.0 ($C_{5/4}$), 115.0 (C_{13}), 105.9 (C_{13}), 71.0 (C_{11}), 67.7 (C_{10}), 45.4 (C_9), 39.5 (C_{19}); **LRMS** (*m/z* +ESI):

[b]Overlap with the C_{19}–**H** peak prevented measurement of the vicinal coupling constant.

found $[M+H]^+$: 444.3, calculated for $C_{21}H_{19}^{35}Cl_2N_5O_2$ $[M+H]^+$: 444.1; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 444.1002. Calculated for $C_{21}H_{20}^{35}Cl_2N_5O_2^+$: 444.0989.

1-((6,7-Dichloro-4-(1-methyl-1H-pyrazol-3-yl)phthalazin-1-yl)amino)-3-phenylpropan-2-ol, 6.23

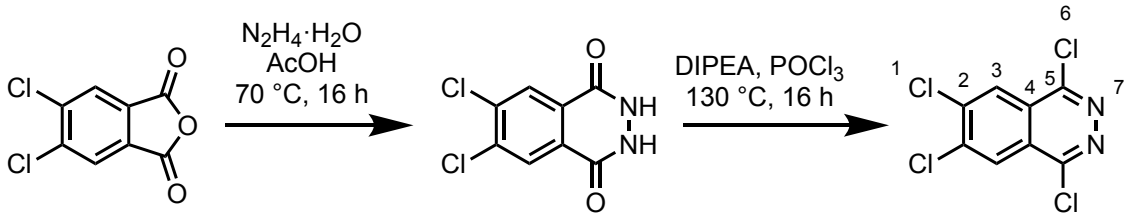


The **title compound** was prepared from 1-phenyl-3-((4,6,7-trichlorophthalazin-1-yl)amino)-propan-2-ol **6.32** (48 mg, 0.12 mmol) and 1-methyl-1*H*-pyrazole-3-boronic acid pinacol ester (31 mg, 0.15 mmol) according to **General Procedure B**. Purification of the crude mixture by flash column chromatography (SiO_2, 50% to 100% EtOAc in cyclohexane) provided the **title compound** as a pale-red amorphous solid (35 mg, 81 µmol, 65% yield).

$\mathbf{R}_f$ (EtOAc) 0.34; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 9.47 (s, 1H, C_6–**H**), 8.73 (s, 1H, C_3–**H**), 7.82 (d, 1H, J = 2.3 Hz, C_{18}–**H**), 7.74 (t, 1H, J = 5.5 Hz, N**H**), 7.31-7.21 (m, 4H, C_{14}–**H** & C_{13}–**H**), 7.16 (tt, 1H, J = 5.2, 3.3 Hz, C_{15}–**H**), 6.87 (d, 1H, J = 2.3 Hz, C_{17}–**H**), 5.12 (d, 1H, J = 4.9 Hz, O**H**), 4.19-4.08 (m, 1H, C_{10}–**H**), 3.99 (s, 3H, C_{19}–**H**), 3.69 (app. dt, 1H, J = 13.0, 5.2 Hz, C_9–**H**), 3.50 (ddd, 1H, J = 13.0, 7.4, 5.5 Hz, C_9'–**H**), 2.86 (dd, 1H, J = 13.7, 4.9 Hz, C_{11}–**H**), 2.72 (dd, 1H, J = 13.7, 7.6 Hz, C_{11}'–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 152.6 (C_8), 150.0 (C_{16}), 142.3 (C_7), 139.9 (C_{12}), 135.0 (C_1), 134.0 (C_2),

132.2 (C_{18}), 129.9 (C_{13}), 128.8 (C_6), 128.4 (C_{14}), 126.2 (C_{15}), 125.3 (C_3), 124.8 ($C_{4/5}$), 118.0 ($C_{5/4}$), 105.9 (C_{17}), 70.0 (C_{10}), 48.2 (C_9), 41.8 (C_{11}), 39.4 (C_{19}); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 428.3, calculated for $C_{21}H_{19}^{35}Cl_2N_5O$ $[M+H]^+$: 428.1; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 428.1026, calculated for $C_{21}H_{20}^{35}Cl_2N_5O^+$: 428.1039.

1,4,6,7-Tetrachlorophthalazine, 6.28

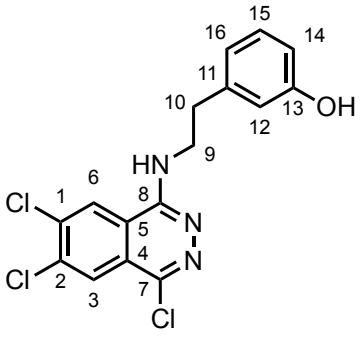


Adapted from a procedure by Lanman *et al.*[398] To a stirred suspension of 5,6-dichloro-1,3-dihydro-2-benzofuran-1,3-dione (20.11 g, 92.68 mmol) in acetic acid (250 mL) was added hydrazine monohydrate (6.0 mL, 120 mmol) dropwise with stirring over the course of one hour at room temperature under an atmosphere of nitrogen. A white solid quickly formed, and the resulting suspension heated under reflux at 70 °C overnight under an atmosphere of nitrogen. The excess acetic acid was removed *in vacuo* and the resultant 6,7-dichloro-2,3-dihydrophthalazine-1,4-dione (21.8 g, 94.2 mmol, quantitative yield) dissolved in phosphorus(V) oxychloride (110 mL, 1.18 mol) without further purification. The resulting suspension was stirred vigorously under an atmosphere of nitrogen until a milk-like dispersion formed, at which point *N,N*-diisopropylethylamine (18.1 mL, 104 mmol) was added dropwise with vigorous stirring, during which a heavy white smoke was evolved and the solution changed from white to yellow. When the addition was complete the resulting yellow solution was heated under reflux at 130 °C overnight, after which the solution took a dark green colour. This was poured carefully whilst still warm into ice-water (500 mL), where it bubbled vio-

lently and adopted a pink colour. The resulting solution was stirred at room temperature for one hour, and the suspension that arose filtered under reduced pressure. The filtrate was extracted with dichloromethane (3 × 30 mL) and the combined organic layers added to the filtrate dissolved in dichloromethane (300 mL). This combined solution was washed with water (3 × 50 mL), brine (50 mL), dried over sodium sulfate, and the solvent removed *in vacuo* to give the crude product as a yellow amorphous solid. Purification by flash column chromatography (SiO_2, 0% to 10% EtOAc in cyclohexane) gave the product as a pale-yellow amorphous solid (6.43 g, 24.0 mmol, 26% yield).

$\mathbf{R}_f$ (70% EtOAc in cyclohexane) 0.78; ^{1}**H NMR** (400 MHz, $CDCl_3$) δ 8.41 (s, 2H, C_2–**H**); ^{13}**C NMR** (100 MHz, $CDCl_3$) δ 153.6 (C_5), 140.4 (C_2), 127.5 (C_3), 126.2 (C_4); **LRMS** (*m/z* +ESI): found $[M+H]^+$, 269.0, calculated for $C_8H_2{}^{35}Cl_4N_2$ $[M+H]^+$: 268.9; **HRMS** (*m/z*): found $[M+H]^+$: 266.9051, calculated for $C_8H_3{}^{35}Cl_4N_2^+$: 266.9045. All data in accordance with literature.[398]

3-(2-((4,6,7-Trichlorophthalazin-1-yl)amino)ethyl)phenol, 6.30

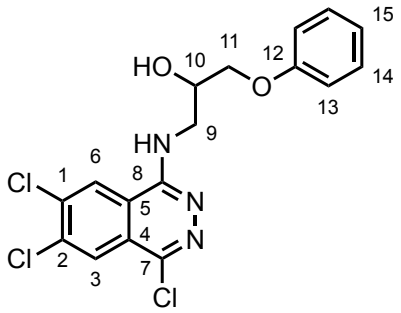


The **title compound** was prepared from 1,4,6,7-tetrachlorophthalazine **6.28** (248 mg, 0.926 mmol) and 3-hydroxyphenethylamine hydrochloride (165 mg, 0.951 mmol) according to **General Procedure A**. Upon completion the reaction was dry-loaded onto silica and pu-

rified by flash column chromatography (SiO_2, 5% to 40% EtOAc in cyclohexane) to give the **title compound** as a pale-yellow amorphous solid (260 mg, 0.705 mmol, 74% yield).

$\mathbf{R}_f$ (40% EtOAc in cyclohexane) 0.36; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 9.27 (br. s, 1H, O**H**), 8.72 (s, 1H, C_6–**H**), 8.22 (s, 1H, C_3–**H**), 7.96 (t, 1H, J = 5.5 Hz, N**H**), 7.08 (t, 1H, J = 7.7 Hz, C_{15}–**H**), 6.71-6.66 (m, 2H, C_{12}–**H** & C_{16}–**H**), 6.60 (ddd, 1H, J = 7.7, 2.4, 1.1 Hz, C_{14}–**H**), 3.74-3.65 (m, 2H, C_9–**H**), 2.90 (dd, 2H, J = 8.5, 6.5 Hz, C_{10}–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 157.4 (C_{13}), 152.8 (C_8), 142.1 ($C_{4/5}$), 141.0 (C_{11}), 135.9 (C_1), 135.4 (C_2), 129.3 (C_{15}), 126.2 (C_6), 125.3 ($C_{4/5}$), 125.0 (C_3), 119.3 (C_{16}), 119.2 (C_7), 115.5 (C_{12}), 113.1 (C_{14}), 42.8 (C_9), 34.1 (C_{10}); **LRMS** (m/z +ESI): found $[M+H]^+$, 371.4, calculated for $C_{16}H_{12}^{35}Cl^{37}Cl_2N_3O$ $[M+H]^+$: 371.0; **HRMS** (m/z): found $[M+H]^+$: 368.0092, calculated for $C_{16}H_{13}^{35}Cl_2N_3O^+$: 368.0119.

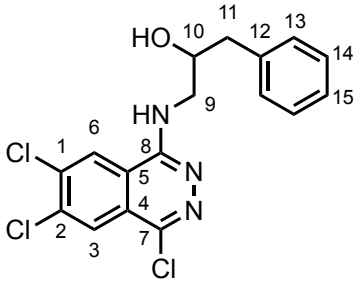
1-Phenoxy-3-((4,6,7-trichlorophthalazin-1-yl)amino)propan-2-ol, 6.31



The **title compound** was prepared from 1,4,6,7-tetrachlorophthalazine **6.28** (311 mg, 1.16 mmol) and 1-amino-3-phenoxypropan-2-ol (204 mg, 1.22 mmol) according to **General Procedure A**. Upon completion the reaction was dry-loaded onto silica and purified by flash column chromatography (SiO_2, 5% to 50% EtOAc in cyclohexane) to provide the **title compound** as a fluffy white solid (315 mg, 0.791 mmol, 65% yield).

$\mathbf{R}_f$ (40% EtOAc in cyclohexane) 0.23; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 8.78 (s, 1H, C_6–**H**), 8.20 (s, 1H, C_3–**H**), 7.94 (t, 1H, J = 5.6 Hz, N**H**), 7.33-7.22 (m, 2H, C_{14}–**H**), 6.98-6.87 (m, 3H, C_{13}–**H** & C_{15}–**H**), 5.34 (br. s, 1H, O**H**), 4.24 (app. p, 1H, J = 5.7 Hz, C_{10}–**H**), 4.04 (dd, 1H, J = 10.0, 4.2 Hz, C_{11}–**H**), 3.98 (dd, 1H, J = 10.0, 5.7 Hz, C_{11}'–**H**), 3.74 (ddd, 1H, J = 13.4, 5.7, 5.6 Hz, C_9–**H**), 3.60 (ddd, 1H, J = 13.1, 7.0, 5.4 Hz, C_9'–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 158.6 (C_{12}), 153.2 (C_8), 142.1 ($C_{4/5}$), 135.9 (C_1), 135.4 (C_2), 129.4 (C_{14}), 126.1 (C_3), 125.6 (C_6), 125.0 ($C_{4/5}$), 120.5 (C_{15}), 119.3 (C_7), 114.5 (C_{13}), 70.4 (C_{11}), 66.8 (C_{10}), 44.8 (C_9); **LRMS** (m/z +ESI): found $[M+H]^+$, 398.1, calculated for $C_{17}H_{14}^{35}Cl_3N_3O_2$ $[M+H]^+$: 398.0; **HRMS** (m/z): found $[M+H]^+$: 398.0239, calculated for $C_{17}H_{15}^{35}Cl_3N_3O_2^+$: 398.0224.

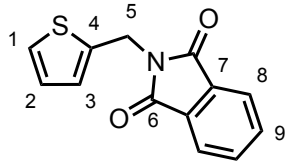
1-Phenyl-3-((4,6,7-trichlorophthalazin-1-yl)amino)propan-2-ol, 6.32



The **title compound** was prepared from 1,4,6,7-tetrachlorophthalazine **6.28** (250 mg, 0.931 mmol) and 1-amino-3-phenyl-propan-2-ol (141 mg, 0.931 mmol) according to **General Procedure A**. Upon completion the reaction was dry-loaded onto silica and purified by flash column chromatography (SiO_2, 5% to 40% EtOAc in cyclohexane) to provide the **title compound** as a pale-yellow amorphous solid (181 mg, 0.473 mmol, 51% yield).

$\mathbf{R}_f$ (40% EtOAc in cyclohexane) 0.31; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 8.74 (s, 1H, C_6–**H**), 8.15 (s, 1H, C_3–**H**), 7.82 (t, 1H, J = 5.5 Hz, N**H**), 7.27-7.22 (m, 4H, C_{13}–**H** & C_{14}–**H**), 7.15 (app. ddd, 1H, J = 8.6, 4.9, 3.9 Hz, C_{15}–**H**), 4.10 (tt, 1H, J = 7.4, 4.7 Hz, C_{10}–**H**), 3.59 (dd, 1H, J = 13.4, 4.7 Hz, C_9–**H**), 3.45 (dd, 1H, J = 13.4, 7.4 Hz, C_9'–**H**), 2.83 (dd, 1H, J = 13.7, 4.7 Hz, C_{11}–**H**), 2.70 (dd, 1H, J = 13.7, 7.4 Hz, C_{11}'–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 153.0 (C_8), 142.0 ($C_{4/5}$), 139.3 (C_{12}), 135.9 (C_1), 135.4 (C_2), 129.4 (C_{13}), 128.0 (C_{14}), 126.1 (C_3), 125.74 (C_6), 125.71 (C_{15}), 125.0 ($C_{4/5}$), 119.3 (C_7), 69.0 (C_{10}), 47.6 (C_9), 41.2 (C_{11}); **LRMS** (*m/z* +ESI): found $[M+H]^+$, 382.2, calculated for $C_{17}H_{14}^{35}Cl_3N_3O$ $[M+H]^+$: 382.0; **HRMS** (*m/z*): found $[M+H]^+$: 382.0293, calculated for $C_{17}H_{15}^{35}Cl_3N_3O^+$: 382.0275.

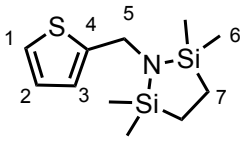
2-(Thiophen-2-ylmethyl)isoindoline-1,3-dione, 6.34



To a solution of phthalic anhydride (1.31 g, 8.82 mmol) in acetic acid (30 mL) was added 2-thiophenemethylamine (1.00 g, 8.82 mmol) dropwise with stirring, and the resulting yellow solution heated under reflux at 130 °C overnight. The resulting yellow-brown solution was cooled to room temperature and water (60 mL) added, causing a white precipitate to form, which was filtered under reduced pressure. The filtrand was washed with water (2 × 10 mL) and dried *in vacuo* to provide the **title compound** as a white amorphous solid, which was carried forward without further purification (1.43 g, 5.88 mmol, 67% yield).

$\mathbf{R}_f$ (40% EtOAc in cyclohexane) 0.73; ^{1}**H NMR** (400 MHz, DMSO−d_6) δ 7.92-7.88 (m, 2H, C_9–**H**), 7.88-7.82 (m, 2H, C_8–**H**), 7.43 (dd, 1H, J = 5.1, 1.3 Hz, C_1–**H**), 7.09 (dd, 1H, J = 3.5, 1.3 Hz, C_3–**H**), 6.96 (dd, 1H, J = 5.1, 3.5 Hz, C_2–**H**), 4.93 (s, 2H, C_5–**H**); ^{13}**C NMR** (100 MHz, DMSO−d_6) δ 167.2 (C_6), 138.6 (C_7), 134.7 (C_9), 131.4 (C_4), 127.03 (C_3), 126.95 (C_2), 126.1 (C_1), 125.3 (C_8), 35.6 (C_5); **HRMS** (*m/z* +ESI): found $[M+H]^+$: 244.0429, calculated for $C_{13}H_{10}NO_2S^+$: 244.0427.

2,2,5,5-Tetramethyl-1-(thiophen-2-ylmethyl)-1,2,5-azadisilolidine, 6.35

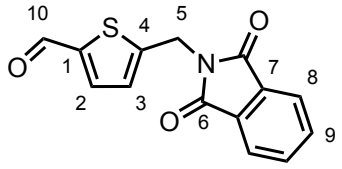


Adapted from the procedure of Muguruma *et al.*[404] To a solution of 2-methylamino thiophene (0.91 mL, 8.8 mmol) in dry dichloromethane (2.5 mL) was added triethylamine (2.5 mL, 18 mmol) dropwise with stirring. A solution of 1,1,4,4-tetramethyl-1,4-dichlorodisilethylene (1.90 g, 8.84 mmol) in dry dichloromethane (10 mL) was added dropwise with stirring, and a white precipitate formed. The resulting suspension was stirred vigorously at room temperature overnight under an atmosphere of dry nitrogen. The white solid was separated by filtration under reduced pressure, and washed with dichloromethane (10 mL), and dried *in vacuo* overnight to provide the **title compound**, which was carried forward without further purification (1.25 g, 4.91 mmol, 56% yield).

^{1}**H NMR** (400 MHz, $CDCl_3$) δ 7.16 (dd, 1H, J = 5.0, 1.4 Hz, C_1–**H**), 6.89 (dd, 1H, J = 5.0, 3.5 Hz, C_2–**H**), 6.87-6.84 (m, 1H, C_3–**H**), 4.18 (d, 2H, J = 0.7 Hz, C_5–**H**), 0.74 (s, 4H, C_7–**H**), 0.03 (s, 12H, C_6–**6**); ^{13}**C NMR** (100 MHz, $CDCl_3$) δ 148.8 (C_4), 126.3 (C_2), 124.1

(C_1 & C_3), 41.1 (C_5), 8.2 (C_7), -0.3 (C_6); *MS data not available due to fragmentation under ESI.* Data in accordance with literature.[404]

5-((1,3-Dioxoisoindolin-2-yl)methyl)thiophene-2-carbaldehyde, 6.37



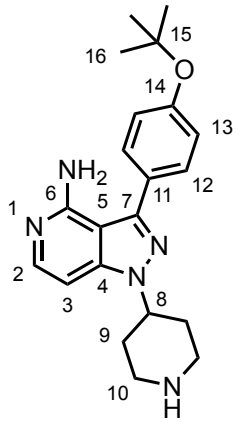
To a stirred solution of dichloromethyl methyl ether (0.29 mL, 3.2 mmol) in dry dichloromethane (25 mL) cooled on ice was added tin(IV) chloride (0.375 mL, 3.21 mmol) and the resulting solution stirred on ice for 20 minutes under an atmosphere of dry nitrogen. 2-(thiophen-2-ylmethyl)isoindoline-1,3-dione **6.34** (600 mg, 2.47 mmol) was added, and the resulting deep-brown solution stirred on ice for 20 minutes before stirring at room temperature under an atmosphere of dry nitrogen overnight. The purple solution was then poured carefully onto ice-water (50 mL) and stirred vigorously at 0 °C for 20 minutes, and at room temperature for 30 minutes. The brown solution was then partitioned, and the aqueous layer washed with dichloromethane (3 × 20 mL), and the combined organic layers washed with water (15 mL), brine (20 mL), and dried in a phase separator. The crude mixture was dry-loaded onto silica and purified by flash column chromatography (SiO_2, 30% to 60% EtOAc in cyclohexane) to provide the **title compound** as a white amorphous solid (478 mg, 1.76 mmol, 71% yield).

^{1}**H NMR** (400 MHz, $CDCl_3$) δ 9.82 (s, 1H, C_{10}–**H**), 7.84 (app. dd, 2H, J = 5.4, 3.0 Hz, C_9–**H**), 7.73 (app. dd, 2H, J = 5.4, 3.0 Hz, C_8–**H**), 7.60 (d, 1H, J = 3.8 Hz, C_2–**H**), 7.20 (d, 1H, J = 3.8 Hz, C_3–**H**) 5.03 (s, 2H, C_5–**H**); ^{13}**C NMR** (100 MHz, $CDCl_3$) δ 182.9 (C_{10}), 167.4 (C_6), 148.6 (C_1), 143.9 (C_4), 136.4 (C_2), 134.4 (C_9), 131.9 (C_7), 128.5 (C_3), 123.7 (C_8), 36.3

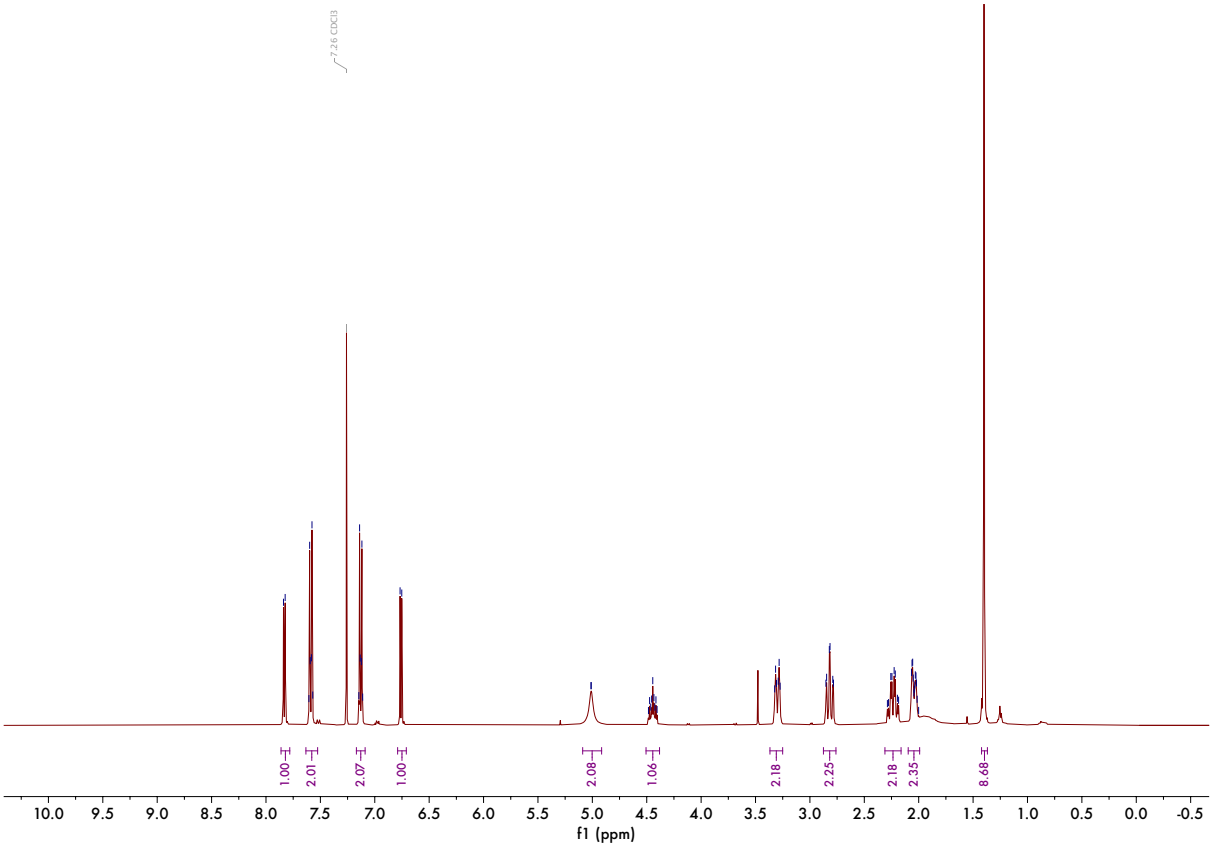
(C_5); **LRMS** (*m/z* +ESI): found $[M+H]^+$: 272.2, calculated for $C_{14}H_9NO_3S$ $[M+H]^+$: 272.0; **HRMS** (*m/z* +ESI): found $[M+H]^+$: 272.0383, calculated for $C_{13}H_{10}NO_2S^+$: 272.0376.

8.5 Selected NMR Spectra

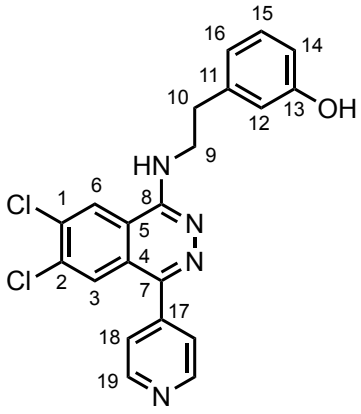
3-(4-(Tert-butoxy)phenyl)-1-(piperidin-4-yl)-1H-pyrazolo[4,3-c]pyridin-4-amine, 6.3



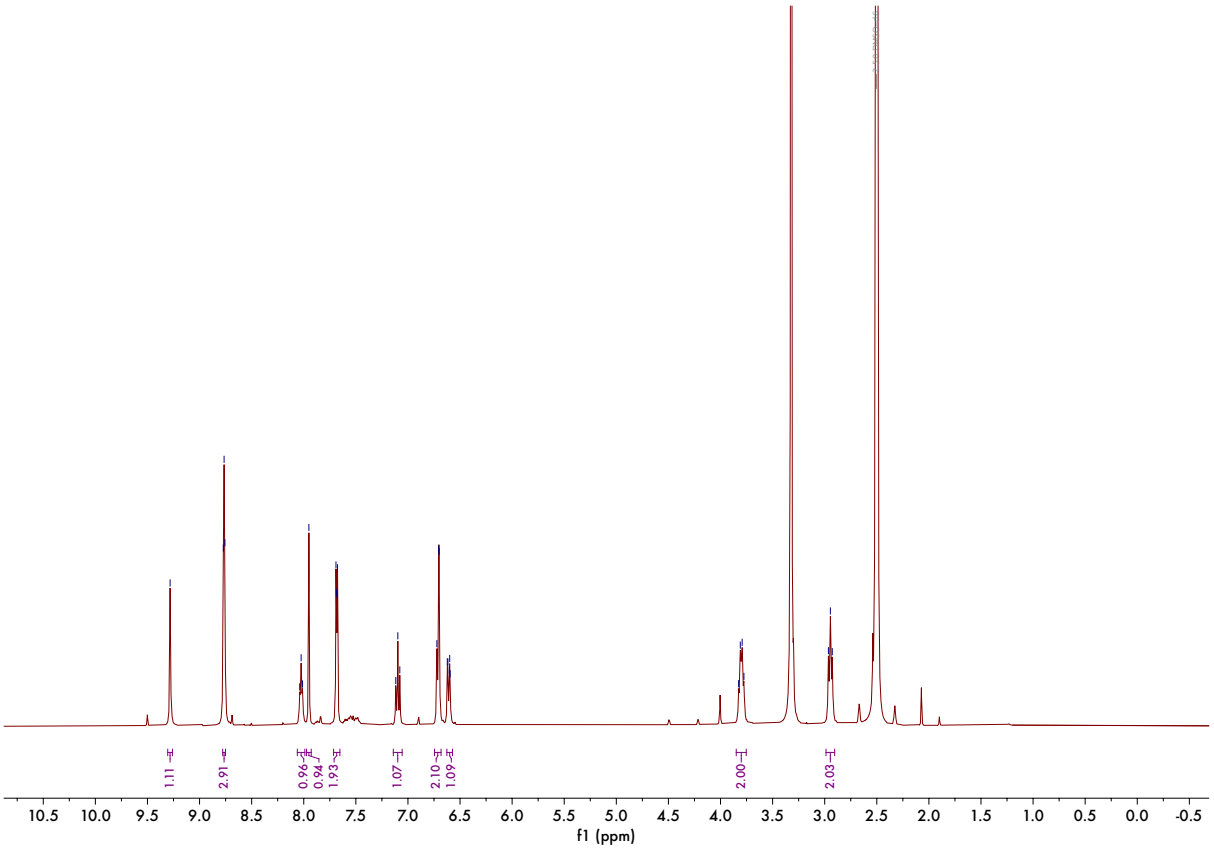
400 MHz ^{1}H NMR spectrum ($CDCl_3$)



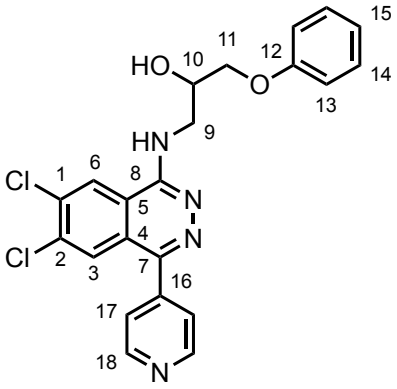
3-(2-((6,7-Dichloro-4-(pyridin-4-yl)phthalazin-1-yl)amino)ethyl)phenol, 6.16



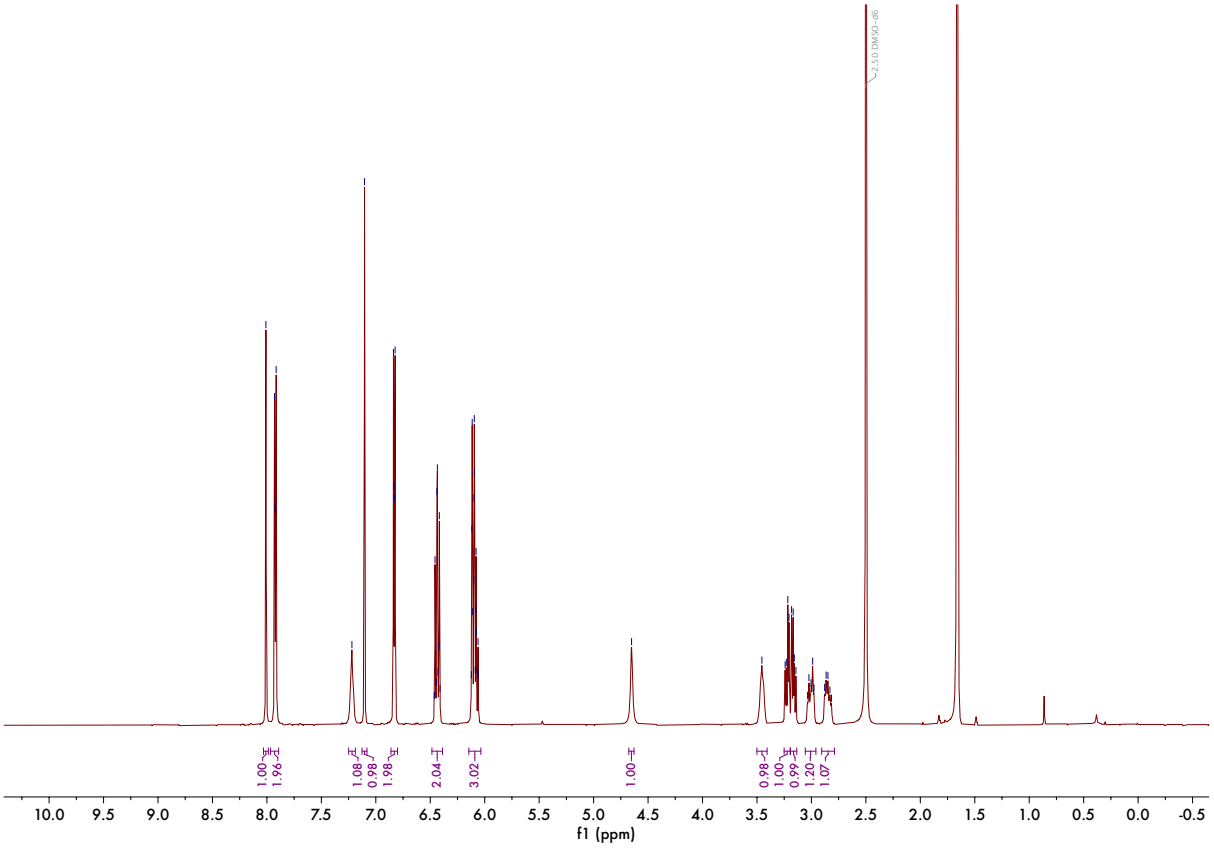
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



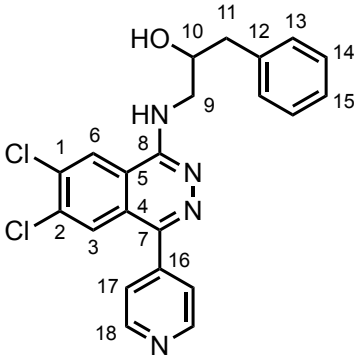
1-((6,7-Dichloro-4-(pyridin-4-yl)phthalazin-1-yl)amino)-3-phenoxypropan-2-ol, 6.17



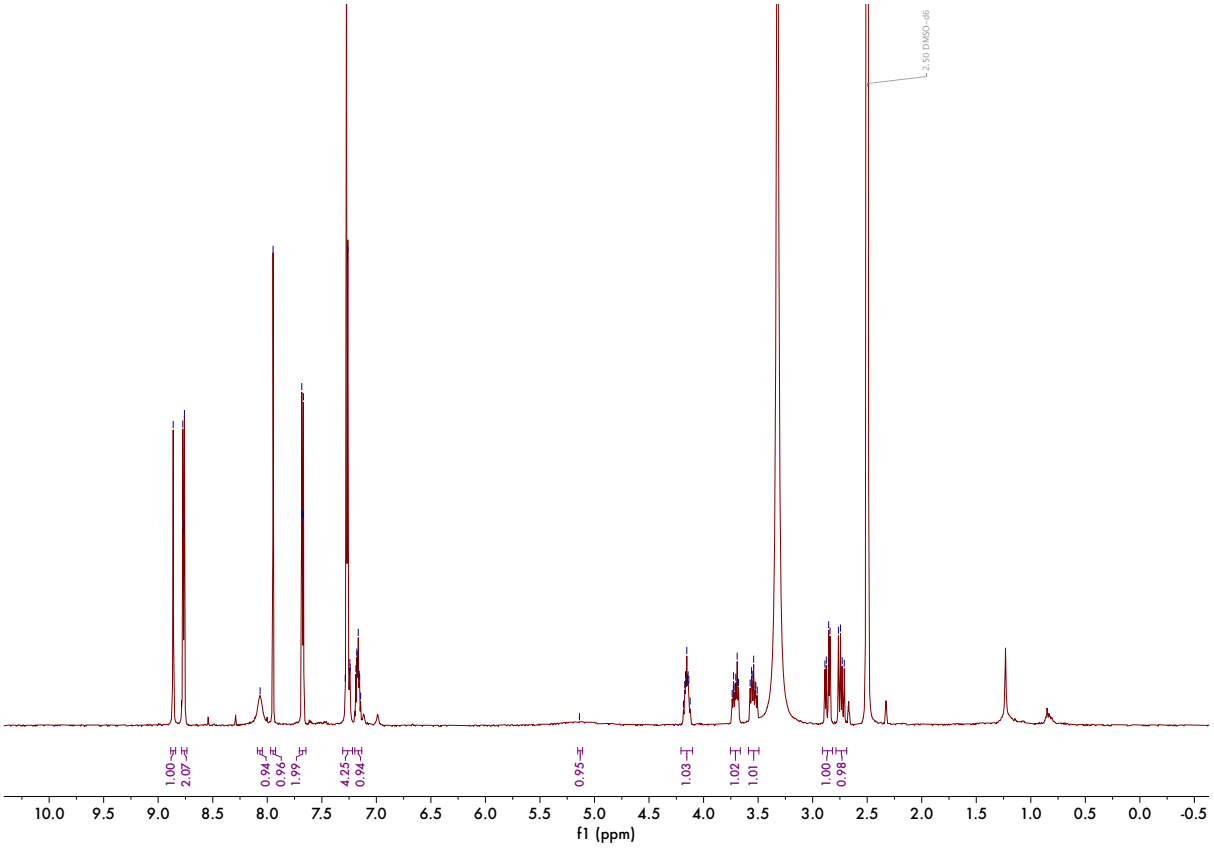
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



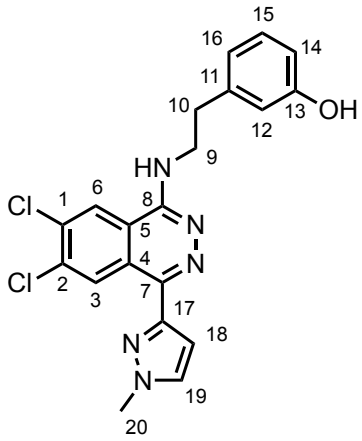
1-((6,7-Dichloro-4-(pyridin-4-yl)phthalazin-1-yl)amino)-3-phenylpropan-2-ol, 6.18



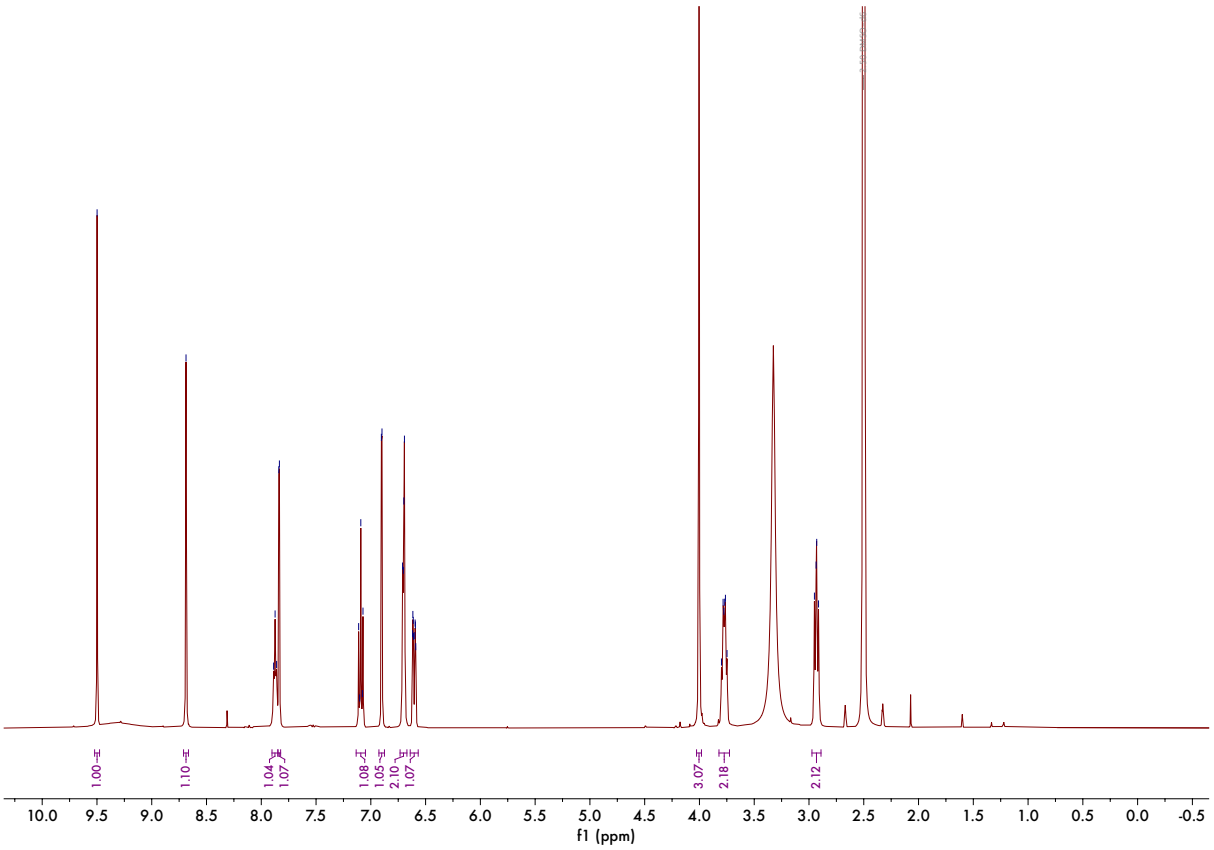
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



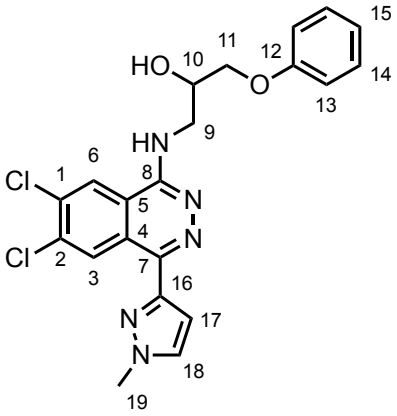
3-(2-((6,7-Dichloro-4-(1-methyl-1H-pyrazol-3-yl)phthalazin-1-yl)amino)ethyl)phenol, 6.21



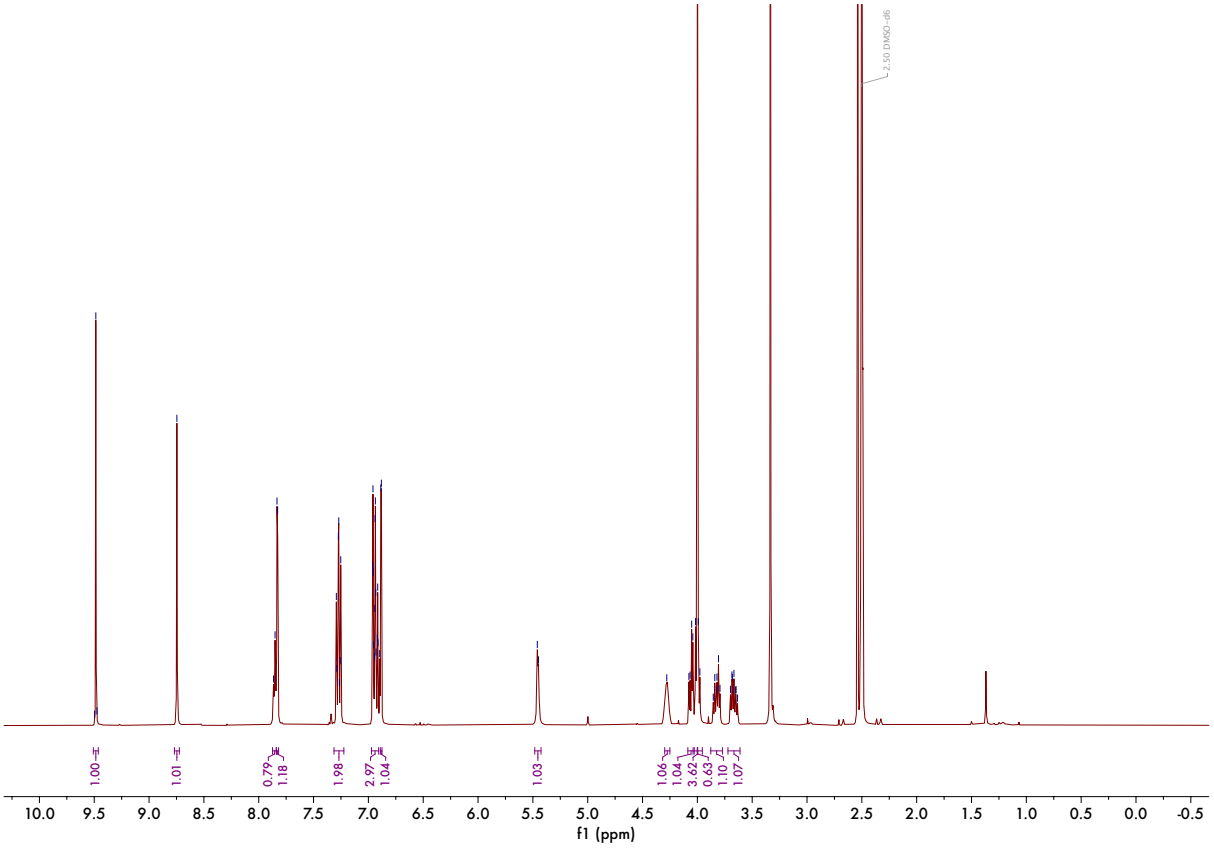
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



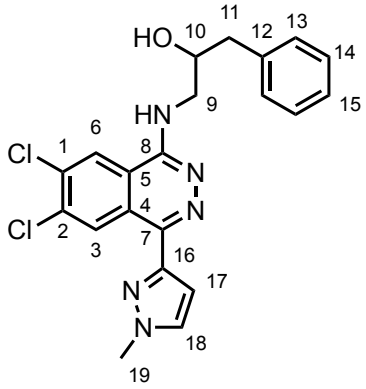
1-((6,7-Dichloro-4-(1-methyl-1H-pyrazol-3-yl)phthalazin-1-yl)amino(-3-phenoxypropan-2-ol, 6.22



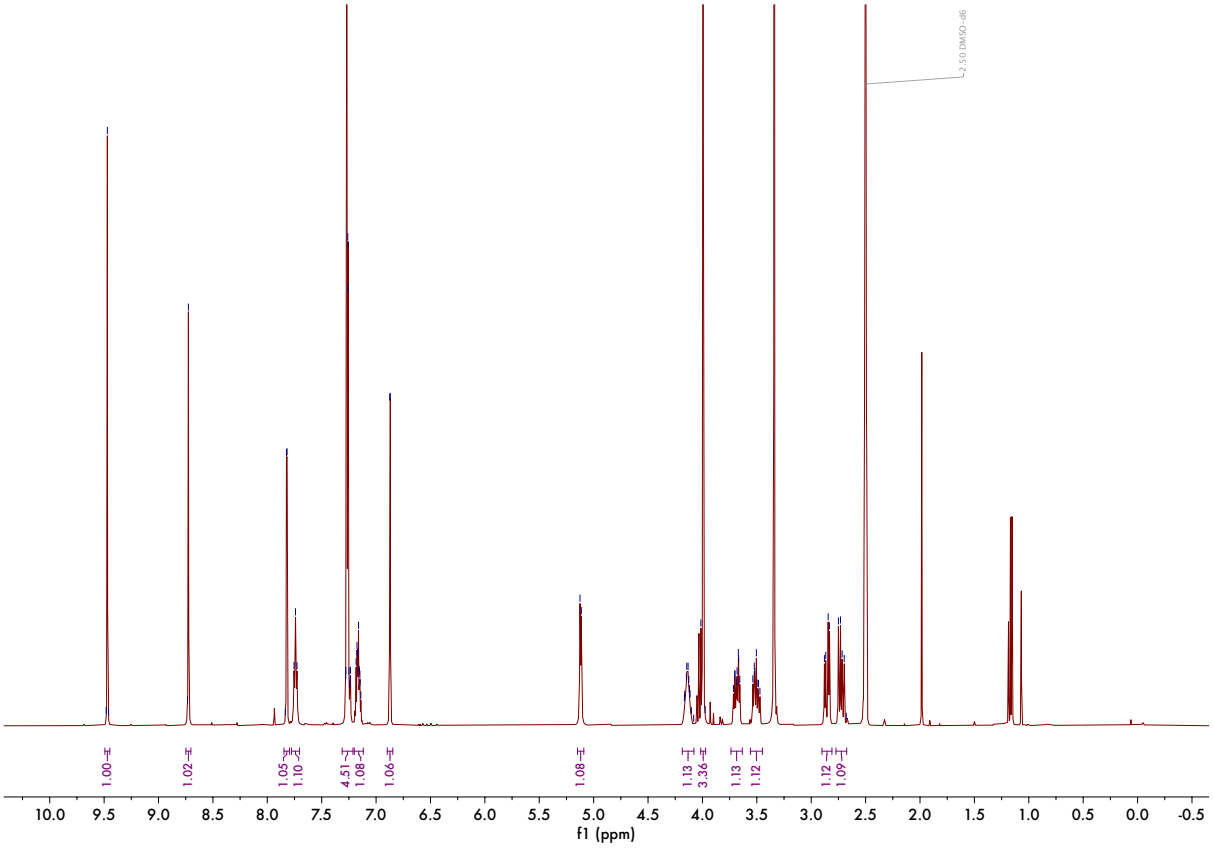
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



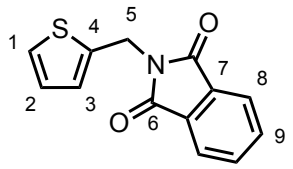
1-((6,7-Dichloro-4-(1-methyl-1H-pyrazol-3-yl)phthalazin-1-yl)amino)-3-phenylpropan-2-ol, 6.23



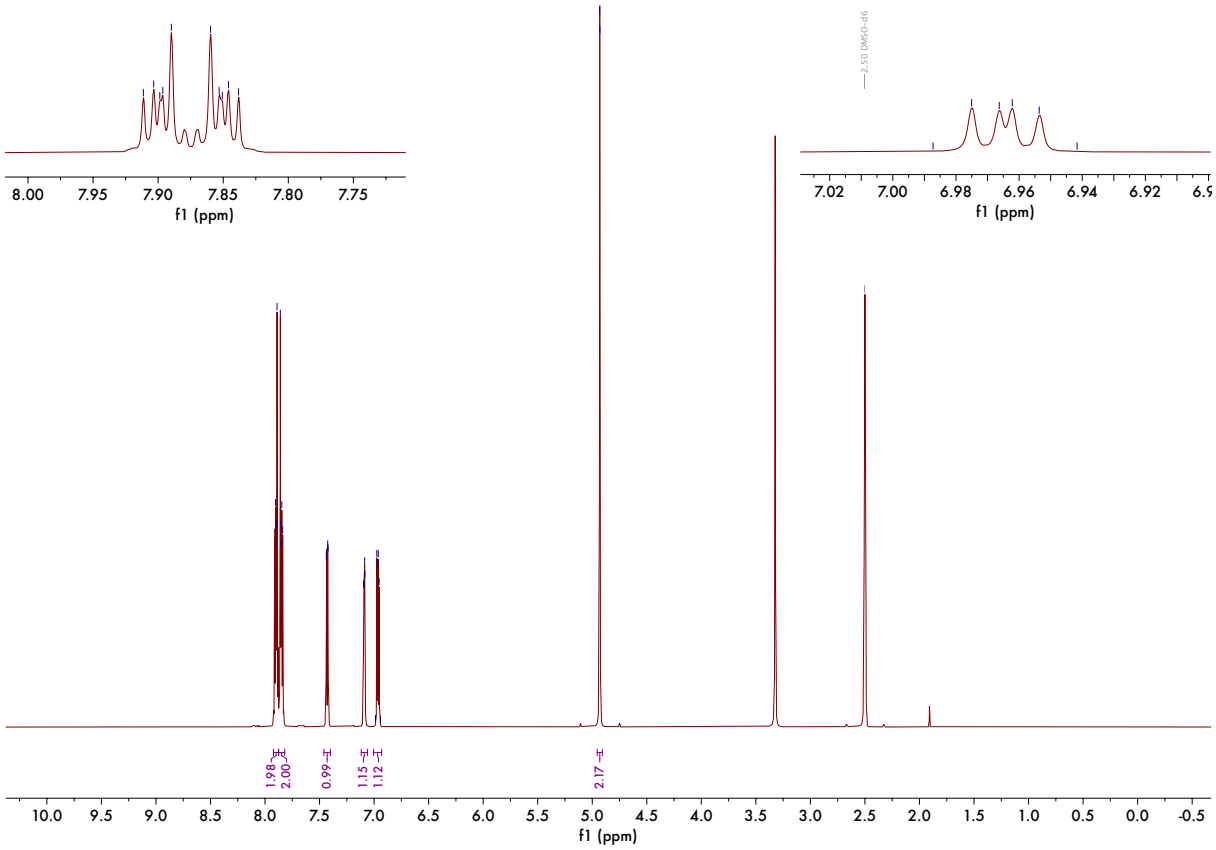
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



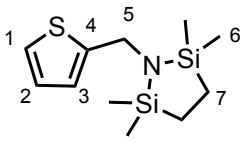
2-(Thiophen-2-ylmethyl)isoindoline-1,3-dione, 6.34



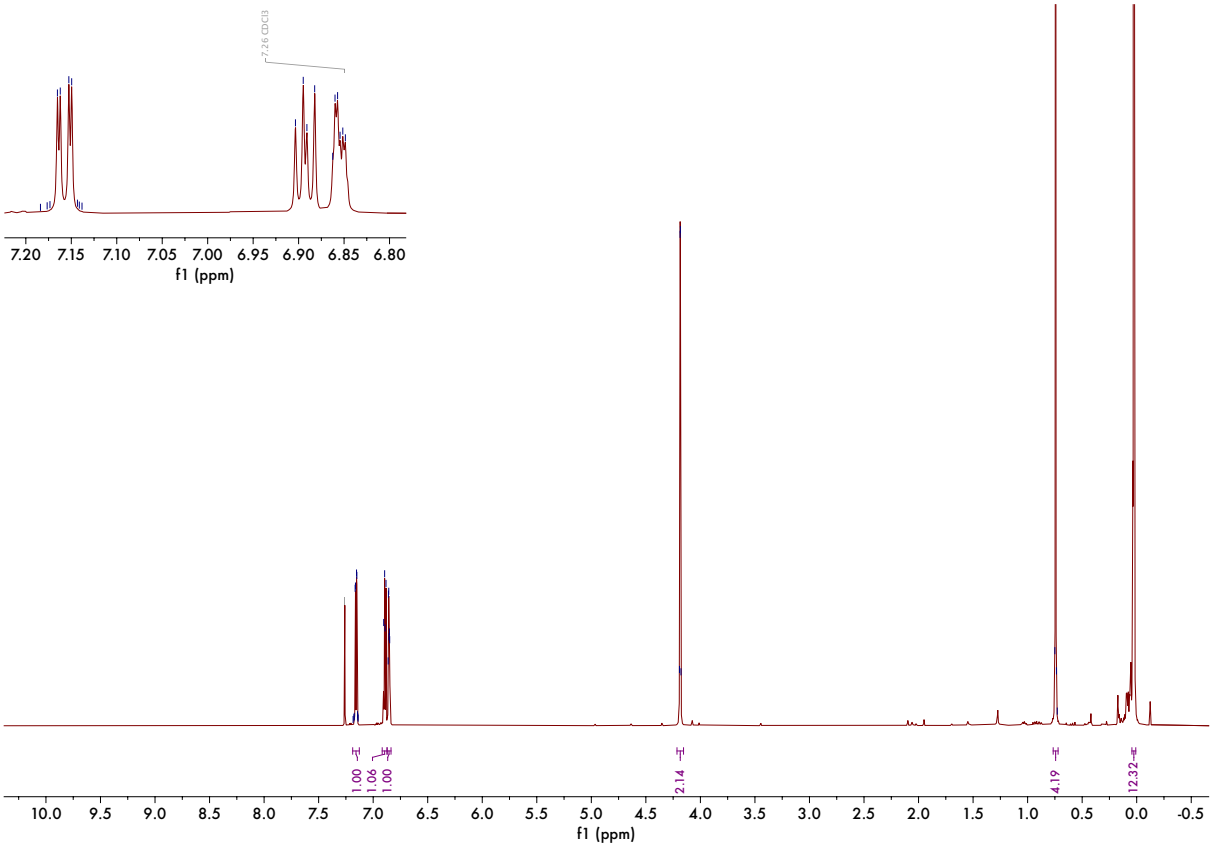
400 MHz ^{1}H NMR spectrum (DMSO−d_6)



2,2,5,5-Tetramethyl-1-(thiophen-2-ylmethyl)-1,2,5-azadisilolidine, 6.35



400 MHz ^{1}H NMR spectrum ($CDCl_3$)



8.6 Biological Assays

8.6.1 NUDT5 and NUDT14 Catalytic Assays

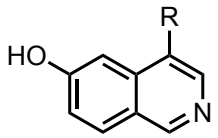
These assays were carried out by Dr Esra Balikçi *and* Dr Klemensas Simelis *at the Centre for Medicines Discovery, University of Oxford.*

Inhibition activities of **6.3** against NUDT5 and NUDT14 were determined using the AMP-Glo™ system (Promega), following a procedure outlined by Balikçi *et al.*[375] The compounds were diluted from 50 µM to 0 µM in a final reaction containing 20 mM HEPES, 100 mM NaCl, 0.5 mM TCEP, 1 mM $MgCl_2$, and 0.1% BSA at pH 7.4. The reactions were performed in 1536-well plates in a 2 µL reaction volume with 1 nM NUDT5 or NUDT14 concentration and 10 µM ADPr as the substrate. The final DMSO concentration was 1% for all reactions. NUDT5 reactions were incubated for 20 minutes, and NUDT14 reactions for one hour, all at room temperature. The reactions were stopped by the addition of 2 µL of AMP-Glo I. The stop solution also contained 25 µM of PubChem CID 16339098 to ensure full arrest of enzyme activity. The reactions were then incubated with 4 µL of the detection solution for one hour at room temperature. Luminescence signals were then measured in a PHERAstar FSX plate reader. Experiments were done in triplicate sets and the data analysed using GraphPad Prism 9 software. Inhibitor dose-response data were normalized to reactions containing vehicle only (1% v/v DMSO, 100% activity) and those containing 500 nM of potent NUDT5 inhibitor TH5427 (1% v/v DMSO, 0% activity).[379] Data are represented as the mean of two independent biological repeats.

9 Appendices

9.1 Excluding X-H Hydrogens from ESP Similarity Calculations

It is helpful to consider an example when validating the exclusion of non-aromatic X-H hydrogens from ESP similarity calculations.



A1

However, when **A1** was searched, the top returned molecule (as shown in Figure 9.1) was not the query. Indeed, **A1** only appears much further down the ranked list of returned molecules, with a rank of 772. Examining the scoring of this revealed that the alignment had a near perfect shape similarity score of 0.99 (it is likely that rounding errors in the calculation process, coupled with very slight differences in atom positioning that are expected in the optimisation process are the reason why this is not exactly 1.00), as expected, but the ESP score was much lower, at 0.45, when the expected value would be 1.00.

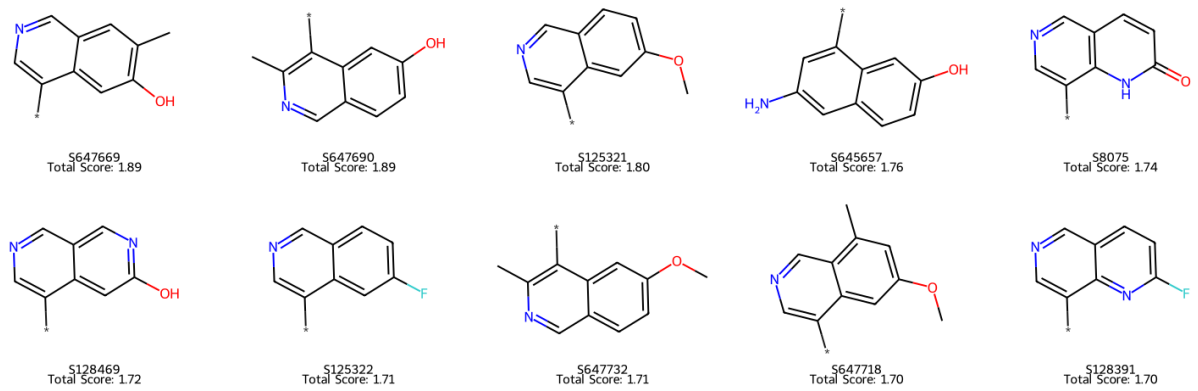


Figure 9.1. The top 10 molecules returned in the HCIE search for **A1**

Visualising the overlay of the returned alignment between query **A1** and the library molecule S125323 (Figure 9.2) showed the alignment of the aromatic atoms and the phenolic oxygen to be perfect, which initially suggested that the unexpectedly low ESP score was not a consequence of a poor alignment, but something more profound. Further inspection of the atom indexing in RDkit after instantiation of the molecules from their respective SMILES strings revealed differences in the atom indexing between the two molecules. This is important as the partial charges are stored in a list structure, where the position in the list of the partial charge is the index of the atom. However, as the atom indexing is common between the coordinates of the atoms and the partial charges, this should not impact the calculation of electrostatic similarity as the similarity (calculated as described in Section 4.2.1.5) depends only on the charges and their respective positions. Inspection of the coordinates of both molecules, and their partial charge distributions as shown in Figure 9.3 verified that, in both molecules, the coordinates and partial charges matched up independent of the atom index.

Observing the alignments shown in Figure 9.2, it is clear that the phenolic hydrogen is not well aligned between the two ligands. When calculating the shape similarity this has no bearing, as only heavy atoms are used in this calculation (as described in Section 4.2.1.5). In-

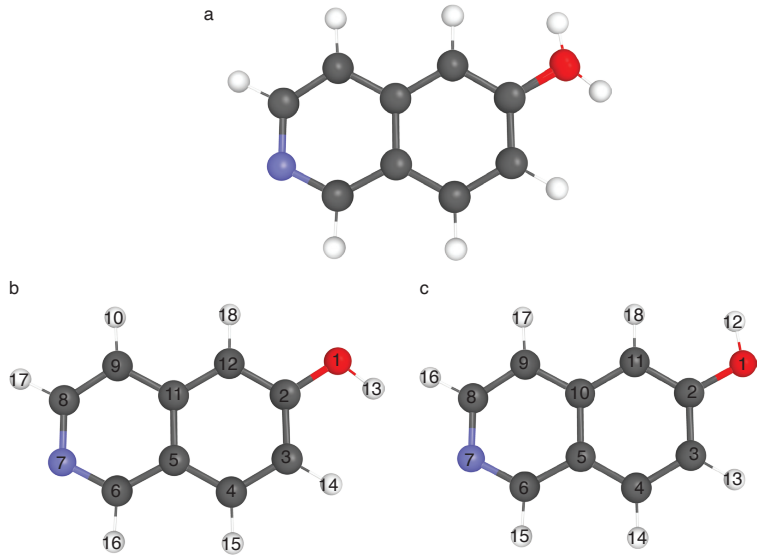
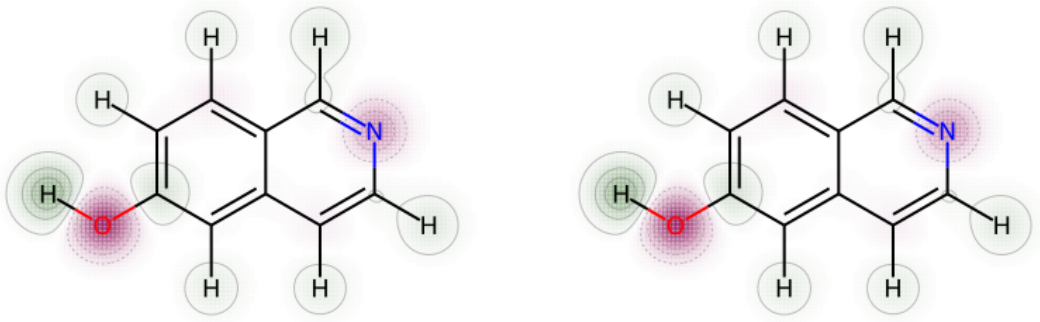


Figure 9.2. a. The alignments of the query molecule **A1** against the library molecule S125323 after HCIE search and alignment. **b.** The RDKit atom indexing of **A1**. **c.** The atom numbering in RDKit of S125323.

specting the partial charge distributions of the two molecules in Figure 9.3 reveals that there is a small but significant $\delta+$ charge on the phenolic hydrogen as a result of the anisotropy of the electron distribution in the oxygen-hydrogen bond. This is located far from the centroid of the molecule, and as such exerts a significant effect on the dipole of the molecule (dipoles being calculated using $\vec{\mu} = \sum \mathrm{Q}\vec{d}$, thus a small partial charge a significant distance from the centre will have a strong effect on the dipole of the molecule). This was verified by calculating the dipole moments of both **A1** and **S125323** (shown in 9.1), which are oriented in significantly different directions, thus explaining the difference in calculated electrostatic similarity.

$$\vec{\mu}_{\mathbf{probe}} = \begin{pmatrix} -0.18 \\ 0.21 \\ -0.05 \end{pmatrix} \quad \vec{\mu}_{\mathbf{A1}} = \begin{pmatrix} 0.09 \\ -0.26 \\ 0.06 \end{pmatrix} \tag{9.1}$$



(a) **(b)**

Figure 9.3. The partial charge distributions of **A1** (**a**) and its library equivalent **S125323**. These molecules have been artificially aligned to aid clarity of comparison, hence why this alignment differs to that shown in Figure 9.2.

To address this, the electrostatic similarity was calculated without reference to hydrogen atoms bonded to non-aromatic atoms. The computational determination of the conformation of the phenolic hydrogen is not a trivial problem, and furthermore the concept of a 'correct' conformation in this context is physically meaningless, as the orientation of these hydrogen-bond donors is likely to be highly variable in solution, and will be very dependent on the binding mode of the ligand when bound. Furthermore, the partial charge on these 'non-aromatic' hydrogens is likely to be consistent enough between the various molecules in the library that discounting them in electrostatic similarity calculations is unlikely to have any effect on the ordering of returned molecules within the results.

This was implemented by the addition of a method on the `Molecule` class, which initially forms a set of the atom indices of the hydrogen atoms bonded to aromatic atoms, and then loops through each atom in the molecule, adding it to a list if the atom is either not a

hydrogen, or has its index in the aforementioned set. A set was chosen over a list due to the increased efficiency of membership checking, with membership checks on a set in Python typically completing in an average time complexity of $O(1)$, compared to a worst-case time complexity of $O(n)$ for lists. The coordinates and charges used to calculate the electrostatic similarity scores in the `calculate_esp_similarity` function are then sliced according to the atom indices in this list.

9.2 VEHICLe Sample for Geometry RMSD Benchmarking

Table 9.1. The RegIDs of the sample of VEHICLe used for geometry RMSD benchmarking.

S3861	S19872	S8957	S21811	S2926	S12961	S18721	S20768
S23901	S269	S7325	S11905	S5244	S767	S13853	S23067
S15914	S17122	S13269	S24766	S24774	S18723	S23006	S5576
S10071	S2748	S18664	S13496	S11926	S12805	S1942	S20022
S22799	S21269	S5335	S18500	S4625	S3951	S21879	S2609
S20367	S14211	S14922	S15018	S17827	S10586	S272	S14542
S23646	S23126	S5679	S395	S23214	S3763	S7165	S10532
S8129	S17526	S13491	S22918	S14447	S9500	S5809	S10719
S14000	S1678	S23208	S24158	S8341	S20336	S18430	S19891
S14772	S1452	S14895	S23145	S14527	S4583	S7993	S5991
S9752	S10704	S3142	S13086	S4547	S1202	S8652	S10309
S1454	S16413	S21147	S632	S4037	S8232	S14753	S1922
S10925	S11426	S23778	S7530	S3017	S20054	S22236	S10445
S23445	S20604	S3140	S17785	S14730	S24030	S23128	S10337
S20170	S254	S1757	S1237	S416	S14871	S18425	S11559
S2767	S7180	S6010	S24213	S22043	S2292	S5149	S7160
S12707	S11787	S16266	S3385	S7885	S22319	S9411	S10135
S7145	S19483	S4405	S19810	S990	S18720	S6736	S20155
S14404	S19454	S15828	S5253	S20522	S15576	S13174	S14541
S12908	S21444	S22935	S16909	S9128	S15238	S21677	S24131
S15524	S9839	S20307	S5290	S20939	S22133	S6939	S16168
S18996	S2987	S18128	S8952	S22874	S406	S20557	S6479
S9368	S14945	S5322	S23224	S5699	S24146	S12687	S5819
S20748	S849	S12377	S22700	S19982	S1162	S10090	S3313
S5817	S22574	S6818	S1369	S16864	S8620	S12446	S19474
S18360	S6869	S4969	S12628	S3982	S7512	S9567	S4681
S697	S20439	S10986	S11645	S22518	S2442	S6126	
S16629	S4501	S17794	S20300	S4579	S20453	S4042	
S4705	S24307	S6398	S17590	S22861	S5107	S10708	
S16833	S13346	S17145	S20894	S3376	S9932	S9002	
S16204	S5019	S10400	S1552	S21367	S2683	S11364	
S13630	S18937	S7333	S22852	S23216	S16283	S9638	

9.3 Bin Boundaries for Two-Vector Geometry Hashing

Table 9.2. The distance bins and their hash code.

Bin (d in Å)	Hash
$0 \leq d \leq 2$	00000
$2.00 \leq d < 2.25$	00001
$2.25 \leq d < 2.50$	00010
$2.50 \leq d < 2.75$	00011
$2.75 \leq d < 3.00$	00100
$3.00 \leq d < 3.25$	00101
$3.25 \leq d < 3.50$	00110
$3.50 \leq d < 3.75$	00111
$3.75 \leq d < 4.00$	01000
$4.00 \leq d < 4.25$	01001
$4.25 \leq d < 4.50$	01010
$4.50 \leq d < 4.75$	01011
$4.75 \leq d < 5.00$	01100
$5.00 \leq d < 5.25$	01101
$5.25 \leq d < 5.50$	01110
$5.50 \leq d < 5.75$	01111
$5.75 \leq d < 6.00$	10000
$6.00 \leq d < \infty$	10001

Table 9.3. The angle bins and their hash code.

Angle	Hash
$0° \leq \alpha_v < 10°$	000
$10° \leq \alpha_v < 25°$	001
$25° \leq \alpha_v < 85°$	010
$85° \leq \alpha_v < 135°$	011
$135° \leq \alpha_v < 165°$	100
$165° \leq \alpha_v \leq 180°$	101

Bibliography

(1) Leroi-Gourhan, A. The flowers found with Shanidar IV, a Neanderthal burial in Iraq. *Science* **1975**, *190*, 562–564, DOI: 10.1126/science.190.4214.562.

(2) Solecki, R. S. The implications of the Shanidar cave Neanderthal flower burial. *Ann. N. Y. Acad. Sci.* **1977**, *293*, 114–124, DOI: 10.1111/j.1749-6632.1977.tb41808.x.

(3) Lietava, J. Medicinal plants in a middle Paleolithic grave Shanidar IV? *J. Ethnopharmacol.* **1992**, *35*, 263–266, DOI: 10.1016/0378-8741(92)90023-k.

(4) Fabricant, D. S.; Farnsworth, N. R. The value of plants used in traditional medicine for drug discovery. *Environ. Health Perspect.* **2001**, *109*, 69–75, DOI: 10.1289/ehp.01109s169.

(5) Yuan, H.; Ma, Q.; Ye, L.; Piao, G. The traditional medicine and modern medicine from natural products. *Molecules* **2016**, *21*, 559, DOI: 10.3390/molecules21050559.

(6) Macht, D. I. A drug or poison? *The Scientific Monthly* **1938**, *47*, 34–40.

(7) Thorn, C. F.; Oshiro, C.; Marsh, S.; Hernandez-Boussard, T.; McLeod, H.; Klein, T. E.; Altman, R. B. Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenet. Genomics* **2011**, *21*, 440–446, DOI: 10.1097/FPC.0b013e32833ffb56.

(8) Gallo, M. A.; Doull, J. History and scope of toxicology, In *Cassarett and Doull's toxicology. The basic science of poisons*, ed. by Klaassen, C. D., McGraw-Hill: 1996, pp 3–11.

(9) Garrison, F. H. F. H., *An introduction to the history of medicine : with medical chronology, suggestions for study and bibliographic data*, 4th ed; Saunders: Philadelphia ; 1929.

(10) Hippocrates, *Hippocrates. Volume VI. Diseases 3. Internal affections. Regimen in acute diseases / with an English translation by W.H.S. Jones and E.T. Withington*; The Loeb Classical Library ; 473. Harvard University Press: Cambridge, MA, 2014.

(11) Sykiotis, G. P.; Kalliolias, G. D.; Papavassiliou, A. G. Pharmacogenetic principles in the Hippocratic writings. *J. Clin. Pharmacol.* **2005**, *45*, 1218–1220, DOI: 10.1177/0091270005281091.

(12) Tsatsakis, A. M.; Vassilopoulou, L; Kovatsi, L; Tsitsimpikou, C; Karamanou, M; Leon, G; Liesivuori, J; Hayes, A. W.; Spandidos, D. A. The dose response principle from philosophy to modern toxicology: The impact of ancient philosophy and medicine in modern toxicology science. *Toxicol. Rep.* **2018**, *5*, 1107–1113, DOI: 10.1016/j.toxrep.2018.10.001.

(13) Li, J.; Corey, E., *Drug discovery: practices, processes, and perspectives*; Wiley: 2013.

(14) Miner, J.; Hoffhines, A. The discovery of aspirin's antithrombotic effects. *Tex. Heart Inst. J.* **2007**, *34*, 179.

(15) Pasipoularides, A. Galen, father of systematic medicine. An essay on the evolution of modern medicine and cardiology. *Int. J. Cardiol.* **2014**, *172*, 47–58, DOI: 10.1016/j.ijcard.2013.12.166.

(16) Borzelleca, J. F. Paracelsus: herald of modern toxicology. *Toxicol. Sci.* **2000**, *53*, 2–4, DOI: 10.1093/toxsci/53.1.2.

(17) Siddiqui, M. A.; Mehta, N. J.; Khan, I. A. Paracelsus: The Hippocrates of the Renaissance. *J. Med. Biogr.* **2003**, *11*, 78–80, DOI: 10.1177/096777200301100207.

(18) Multhauf, R. Medical chemistry and "The Paracelsians". *Bull. Hist. Med.* **1954**, *28*, 101–126.

(19) Deichmann, W. B.; Henschler, D; Holmstedt, B; Keil, G What is there that is not poison? A study of the Third Defense by Paracelsus. *Arch. Toxicol.* **1986**, *58*, 207–213, DOI: 10.1007/BF00297107.

(20) Lockermann, G. Friedrich Wilhelm Serturner, the discoverer of morphine. *J. Chem. Educ.* **1951**, *28*, 277, DOI: 10.1021/ed028p277.

(21) Schmitz, R. Friedrich Wilhelm Sertürner and the discovery of morphine. *Pharm. Hist.* **1985**, *27*, 61–74.

(22) Chast, F. A history of drug discovery: From first steps of chemistry to achievements in molecular pharmacology, In *The Practice of Medicinal Chemistry*, ed. by Wermuth, C. G., Academic Press: 2008, pp 1–62, DOI: 10.1016/B978-0-12-374194-3.00001-9.

(23) Sneader, W The discovery of aspirin: a reappraisal. *BMJ* **2000**, *321*, 1591–1594, DOI: 10.1136/bmj.321.7276.1591.

(24) Desborough, M. J. R.; Keeling, D. M. The aspirin story - from willow to wonder drug. *Br. J. Haematol.* **2017**, *177*, 674–683, DOI: 10.1111/bjh.14520.

(25) Pasteur, L.; Chamberland, C.; Joubert, J., et al. Théorie des germes et ses applications à la médecine et à la chirurgie. 1878.

(26) Bynum, W. F., *The History of Medicine: A Very Short Introduction*; Very Short Introductions; Oxford University Press: London, England, 2008.

(27) Hutchings, M. I.; Truman, A. W.; Wilkinson, B. Antibiotics: past, present and future. *Curr. Opin. Microbiol.* **2019**, *51*, 72–80, DOI: 10.1016/j.mib.2019.10.008.

(28) Lehrer, S., *Explorers of the Body*, 2nd ed.; iUniverse: 2006.

(29) The Nobel Prize in Physiology or Medicine 1945 nobelprize.org, `https://www.nobelprize.org/prizes/medicine/1945/summary/`, [Accessed 04-01-2025].

(30) Hodgkin, D. C. The X-ray analysis of the structure of penicillin. *Adv. Sci.* **1949**, *6*, 85–89.

(31) Hochhaus, G; Barrett, J. S.; Derendorf, H Evolution of pharmacokinetics and pharmacokinetic/dynamic correlations during the 20th century. *J. Clin. Pharmacol.* **2000**, *40*, 908–917, DOI: `10.1177/00912700022009648`.

(32) Pereira, D. A.; Williams, J. A. Origin and evolution of high throughput screening. *Br. J. Pharmacol.* **2007**, *152*, 53–61, DOI: `10.1038/sj.bjp.0707373`.

(33) Rotella, D. P. The critical role of organic chemistry in drug discovery. *ACS Chem. Neurosci.* **2016**, *7*, 1315–1316, DOI: `10.1021/acschemneuro.6b00280`.

(34) Roberts, N. A.; Martin, J. A.; Kinchington, D; Broadhurst, A. V.; Craig, J. C.; Duncan, I. B.; Galpin, S. A.; Handa, B. K.; Kay, J; Kröhn, A Rational design of peptide-based HIV proteinase inhibitors. *Science* **1990**, *248*, 358–361, DOI: `10.1126/science.2183354`.

(35) Erickson, J; Neidhart, D. J.; VanDrie, J; Kempf, D. J.; Wang, X. C.; Norbeck, D. W.; Plattner, J. J.; Rittenhouse, J. W.; Turon, M; Wideburg, N Design, activity, and 2.8 Å crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease. *Science* **1990**, *249*, 527–533, DOI: `10.1126/science.2200122`.

(36) Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787–797, DOI: `10.1016/j.chembiol.2003.09.002`.

(37) Hantschel, O. Unexpected off-targets and paradoxical pathway activation by kinase inhibitors. *ACS Chem. Biol.* **2015**, *10*, 234–245, DOI: `10.1021/cb500886n`.

(38) Cohen, P.; Cross, D.; Jänne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat. Rev. Drug Discov.* **2021**, *20*, 551–569, DOI: 10.1038/s41573-021-00195-4.

(39) Kantarjian, H.; Jabbour, E.; O'Brien, S. Chronic myelogenous leukemia, In *Molecular Hematology*, Wiley: 2024, pp 83–97, DOI: 10.1002/9781119252863.ch6.

(40) Shaker, B.; Ahmad, S.; Lee, J.; Jung, C.; Na, D. In silico methods and tools for drug discovery. *Comput. Biol. Med.* **2021**, *137*, 104851, DOI: 10.1016/j.compbiomed.2021.104851.

(41) Gund, P; Andose, J. D.; Rhodes, J. B.; Smith, G. M. Three-dimensional molecular modeling and drug design. *Science* **1980**, *208*, 1425–1431, DOI: 10.1126/science.6104357.

(42) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288, DOI: 10.1016/0022-2836(82)90153-x.

(43) Hartman, G. D.; Egbertson, M. S.; Halczenko, W; Laswell, W. L.; Duggan, M. E.; Smith, R. L.; Naylor, A. M.; Manno, P. D.; Lynch, R. J.; Zhang, G Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *J. Med. Chem.* **1992**, *35*, 4640–4642, DOI: 10.1021/jm00102a020.

(44) Van Drie, J. H. Computer-aided drug design: the next 20 years. *J. Comput. Aided Mol. Des.* **2007**, *21*, 591–601, DOI: 10.1007/s10822-007-9142-y.

(45) Talele, T.; Khedkar, S.; Rigby, A. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* **2010**, *10*, 127–141, DOI: 10.2174/156802610790232251.

(46) Sadybekov, A. V.; Katritch, V. Computational approaches streamlining drug discovery. *Nature* **2023**, *616*, 673–685, DOI: 10.1038/s41586-023-05905-z.

(47) Research and development in the pharmaceutical industry -cbo.gov, https://www.cbo.gov/publication/57126, [Accessed 04-01-2025].

(48) Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B.* **2022**, *12*, 3049–3062, DOI: 10.1016/j.apsb.2022.02.002.

(49) The Nobel Prize in Chemistry 2024, https://www.nobelprize.org/prizes/chemistry/2024/summary/, [Accessed 25-11-2024].

(50) Martins, A. C.; Oshiro, M. Y.; Albericio, F.; de la Torre, B. G. Food and Drug Administration (FDA) approvals of biological drugs in 2023. *Biomedicines* **2024**, *12*, 1992, DOI: 10.3390/biomedicines12091992.

(51) G de la Torre, B.; Albericio, F. The pharmaceutical industry in 2018. An analysis of FDA drug approvals from the perspective of molecules. *Molecules* **2019**, *24*, 809, DOI: 10.3390/molecules24040809.

(52) Mullard, A. 2019 FDA drug approvals. *Nat. Rev. Drug Discov.* **2020**, *19*, 79–84, DOI: 10.1038/d41573-020-00001-7.

(53) Mullard, A. 2020 FDA drug approvals. *Nat. Rev. Drug Discov.* **2021**, *20*, 85–90, DOI: 10.1038/d41573-021-00002-0.

(54) Mullard, A. 2021 FDA approvals. *Nat. Rev. Drug Discov.* **2022**, *21*, 83–88, DOI: 10.1038/d41573-022-00001-9.

(55) Mullard, A. 2022 FDA approvals. *Nat. Rev. Drug Discov.* **2023**, *22*, 83–88, DOI: 10.1038/d41573-023-00001-3.

(56) Mullard, A. 2023 FDA approvals. *Nat. Rev. Drug Discov.* **2024**, *23*, 88–95, DOI: 10.1038/d41573-024-00001-x.

(57) Mullard, A. 2024 FDA approvals. *Nat. Rev. Drug Discov.* **2025**, *24*, 75–82, DOI: 10.1038/d41573-025-00001-5.

(58) Vitaku, E.; Smith, D. T.; Njardarson, J. T. Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals. *J. Med. Chem.* **2014**, *57*, 10257–10274, DOI: 10.1021/jm501100b.

(59) Vallianatou, T.; Giaginis, C.; Tsantili-Kakoulidou, A. The impact of physicochemical and molecular properties in drug design: Navigation in the "drug-like" chemical space, In *Advances in Experimental Medicine and Biology*, Springer International Publishing: Cham, 2015, pp 187–194, DOI: 10.1007/978-3-319-08927-0_21.

(60) Beck, H.; Härter, M.; Haß, B.; Schmeck, C.; Baerfacker, L. Small molecules and their impact in drug discovery: A perspective on the occasion of the 125th anniversary of the Bayer Chemical Research Laboratory. *Drug Discov. Today* **2022**, *27*, 1560–1574, DOI: 10.1016/j.drudis.2022.02.015.

(61) Xu, X.; Vugmeyster, Y. Challenges and opportunities in absorption, distribution, metabolism, and excretion studies of therapeutic biologics. *AAPS J.* **2012**, *14*, 781–791, DOI: 10.1208/s12248-012-9388-8.

(62) Sathish, J. G. et al. Challenges and approaches for the development of safer immunomodulatory biologics. *Nat. Rev. Drug Discov.* **2013**, *12*, 306–324, DOI: 10.1038/nrd3974.

(63) Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, DOI: 10.1038/srep42717.

(64) Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **2014**, *32*, 40–51, DOI: 10.1038/nbt.2786.

(65) Southey, M. W. Y.; Brunavs, M. Introduction to small molecule drug discovery and preclinical development. *Front. Drug Discov.* **2023**, *3*, DOI: 10.3389/fddsv.2023.1314077.

(66) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26, DOI: 10.1016/S0169-409X(96)00423-1.

(67) Ouyang, J.; Zhang, Z.; Deng, B.; Liu, J.; Wang, L.; Liu, H.; Koo, S.; Chen, S.; Li, Y.; Yaremenko, A. V.; Huang, X.; Chen, W.; Lee, Y.; Tao, W. Oral drug delivery platforms for biomedical applications. *Mater. Today* **2023**, *62*, 296–326, DOI: 10.1016/j.mattod.2023.01.002.

(68) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337–341, DOI: 10.1016/j.ddtec.2004.11.007.

(69) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **2007**, *6*, 881–890, DOI: 10.1038/nrd2445.

(70) Schreiber, S. L. Organic synthesis toward small-molecule probes and drugs. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6699–6702, DOI: 10.1073/pnas.1103205108.

(71) Galloway, W. R. J. D.; Isidro-Llobet, A.; Spring, D. R. Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun.* **2010**, *1*, 80, DOI: 10.1038/ncomms1081.

(72) Ottosson, H. A focus on aromaticity: fuzzier than ever before? *Chem. Sci.* **2023**, *14*, 5542–5544, DOI: 10.1039/d3sc90075d.

(73) Merino, G.; Solà, M.; Fernández, I.; Foroutan-Nejad, C.; Lazzeretti, P.; Frenking, G.; Anderson, H. L.; Sundholm, D.; Cossío, F. P.; Petrukhina, M. A.; Wu, J.; Wu,

J. I.; Restrepo, A. Aromaticity: quo vadis. *Chem. Sci.* **2023**, *14*, 5569–5576, DOI: 10.1039/D2SC04998H.

(74) Balaban, A. T.; Oniciu, D. C.; Katritzky, A. R. Aromaticity as a Cornerstone of Heterocyclic Chemistry. *Chem. Rev.* **2004**, *104*, 2777–2812, DOI: 10.1021/cr0306790.

(75) Joule, J. A.; Mills, K., *Heterocyclic Chemistry*, 5th ed.; Wiley-Blackwell: Chichester, England, 2010.

(76) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. Solvent effects. 3. Tautomeric equilibria of formamide and 2-pyridone in the gas phase and solution: an ab initio SCRF study. *J. Am. Chem. Soc.* **1992**, *114*, 1645–1652, DOI: 10.1021/ja00031a017.

(77) Gadosy, T. A.; McClelland, R. A. The keto/enol equilibrium of phenol. An ab initio investigation. *Comput. Theor. Chem.* **1996**, *369*, 1–8, DOI: 10.1016/S0166-1280(96)04694-5.

(78) Bielefeld, M. J.; Fitts, D. D. Bonding Contribution of Sulfur d Orbitals in Thiophene. An Extension of the Self-Consistent Field Molecular Orbital Method1. *J. Am. Chem. Soc.* **1966**, *88*, 4804–4810, DOI: 10.1021/ja00973a008.

(79) Marshall, C. M.; Federice, J. G.; Bell, C. N.; Cox, P. B.; Njardarson, J. T. An update on the nitrogen heterocycle compositions and properties of US FDA-approved pharmaceuticals (2013-2023). *J. Med. Chem.* **2024**, *67*, 11622–11655, DOI: 10.1021/acs.jmedchem.4c01122.

(80) Delost, M. D.; Smith, D. T.; Anderson, B. J.; Njardarson, J. T. From oxiranes to oligomers: Architectures of U.S. FDA-approved pharmaceuticals containing oxygen heterocycles. *J. Med. Chem.* **2018**, *61*, 10996–11020, DOI: 10.1021/acs.jmedchem.8b00876.

(81) Meanwell, N. Chapter five - a synopsis of the properties and applications of heteroaromatic rings in medicinal chemistry, In *Advances in Heterocyclic Chemistry*,

ed. by Scriven, E. F.; Ramsden, C. A., Academic Press: 2017, pp 245–361, DOI: 10.1016/bs.aihch.2016.11.002.

(82) Heravi, M. M.; Zadsirjan, V. Prescribed drugs containing nitrogen heterocycles: an overview. *RSC Adv.* **2020**, *10*, 44247–44311, DOI: 10.1039/D0RA09198G.

(83) Hajduk, P. J.; Galloway, W. R. J. D.; Spring, D. R. Drug discovery: A question of library design. *Nature* **2011**, *470*, 42–43, DOI: 10.1038/470042a.

(84) Lazzara, P. R.; Moore, T. W. Scaffold-hopping as a strategy to address metabolic liabilities of aromatic compounds. *RSC Med. Chem.* **2020**, *11*, 18–29, DOI: 10.1039/C9MD00396G.

(85) Alkorta, I; Elguero, J How aromaticity affects the chemical and physicochemical properties of heterocycles: A computational approach, In *Topics in Heterocyclic Chemistry*, Springer Berlin Heidelberg: Berlin, Heidelberg, 2008, pp 155–202, DOI: 10.1007/978-3-540-68343-8_4.

(86) Nittinger, E.; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K. T.; Lange, G.; Klein, R.; Rarey, M. Large-scale analysis of hydrogen bond interaction patterns in protein-ligand interfaces. *J. Med. Chem.* **2017**, *60*, 4245–4257, DOI: 10.1021/acs.jmedchem.7b00101.

(87) Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084, DOI: 10.1021/jm100112j.

(88) Liu, Z.; Wang, G.; Li, Z.; Wang, R. Geometrical preferences of the hydrogen bonds on protein-ligand binding interface derived from statistical surveys and quantum mechanics calculations. *J. Chem. Theory Comput.* **2008**, *4*, 1959–1973, DOI: 10.1021/ct800267x.

(89) Sharma, V.; Gupta, M. Designing of kinase hinge binders: A medicinal chemistry perspective. *Chem. Biol. Drug. Des.* **2022**, *100*, 968–980, DOI: 10.1111/cbdd.14024.

(90) Wang, B.; Wu, H.; Hu, C.; Wang, H.; Liu, J.; Wang, W.; Liu, Q. An overview of kinase downregulators and recent advances in discovery approaches. *Signal Transduct. Target. Ther.* **2021**, *6*, 423, DOI: 10.1038/s41392-021-00826-7.

(91) Xing, L.; Klug-Mcleod, J.; Rai, B.; Lunney, E. A. Kinase hinge binding scaffolds and their hydrogen bond patterns. *Bioorg. Med. Chem.* **2015**, *23*, 6520–6527, DOI: 10.1016/j.bmc.2015.08.006.

(92) Noble, M. E. M.; Endicott, J. A.; Johnson, L. N. Protein kinase inhibitors: insights into drug design from structure. *Science* **2004**, *303*, 1800–1805, DOI: 10.1126/science.1095920.

(93) Faber, E. B. et al. Development of allosteric and selective CDK2 inhibitors for contraception with negative cooperativity to cyclin binding. *Nat. Commun.* **2023**, *14*, 3213, DOI: 10.1038/s41467-023-38732-x.

(94) Wilkinson, R. D. A.; Williams, R.; Scott, C. J.; Burden, R. E. Cathepsin S: therapeutic, diagnostic, and prognostic potential. *Biol. Chem.* **2015**, *396*, 867–882, DOI: 10.1515/hsz-2015-0114.

(95) Asaad, N. et al. Dipeptidyl nitrile inhibitors of Cathepsin L. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4280–4283, DOI: 10.1016/j.bmcl.2009.05.071.

(96) Bethel, P. A.; Gerhardt, S.; Jones, E. V.; Kenny, P. W.; Karoutchi, G. I.; Morley, A. D.; Oldham, K.; Rankine, N.; Augustin, M.; Krapp, S.; Simader, H.; Steinbacher, S. Design of selective Cathepsin inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4622–4625, DOI: 10.1016/j.bmcl.2009.06.090.

(97) Meanwell, N. A. Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety. *Chem. Res. Toxicol.* **2011**, *24*, 1420–1456, DOI: 10.1021/tx200211v.

(98) Johnson, T. W.; Gallego, R. A.; Edwards, M. P. Lipophilic efficiency as an important metric in drug design. *J. Med. Chem.* **2018**, *61*, 6401–6420, DOI: 10.1021/acs.jmedchem.8b00077.

(99) Ishikawa, M.; Hashimoto, Y. Improvement in aqueous solubility in small molecule drug discovery programs by disruption of molecular planarity and symmetry. *J. Med. Chem.* **2011**, *54*, 1539–1554, DOI: 10.1021/jm101356p.

(100) Van de Waterbeemd, H.; Smith, D. A.; Jones, B. C. Lipophilicity in PK design: methyl, ethyl, futile. *J. Comput. Aided Mol. Des.* **2001**, *15*, 273–286, DOI: 10.1023/a:1008192010023.

(101) Bhal, S. K.; Kassam, K.; Peirson, I. G.; Pearl, G. M. The rule of five revisited: applying log D in place of log P in drug-likeness filters. *Mol. Pharm.* **2007**, *4*, 556–560, DOI: 10.1021/mp0700209.

(102) Lobo, S. Is there enough focus on lipophilicity in drug discovery? *Expert Opin. Drug Discov.* **2020**, *15*, 261–263, DOI: 10.1080/17460441.2020.1691995.

(103) Hill, A. P.; Young, R. J. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discov. Today* **2010**, *15*, 648–655, DOI: 10.1016/j.drudis.2010.05.016.

(104) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250–1256, DOI: 10.1021/jm021053p.

(105) Leeson, P. D.; St-Gallay, S. A. The influence of the 'organizational factor' on compound quality in drug discovery. *Nat. Rev. Drug Discov.* **2011**, *10*, 749–765, DOI: 10.1038/nrd3552.

(106) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51*, 817–834, DOI: 10.1021/jm701122q.

(107) Jampilek, J. Heterocycles in medicinal chemistry. *Molecules* **2019**, *24*, 3839, DOI: 10.3390/molecules24213839.

(108) Wood, A.; Armour, D. The discovery of the CCR5 receptor antagonist, UK-427,857, A new agent for the treatment of HIV infection and AIDS, In *Progress in Medicinal Chemistry*, ed. by King, F.; Lawton, G., Elsevier: 2005, pp 239–271, DOI: 10.1016/S0079-6468(05)43007-6.

(109) Stupple, P. A. et al. An imidazopiperidine series of CCR5 antagonists for the treatment of HIV: the discovery of N-{(1S)-1-(3-fluorophenyl)-3-[(3-endo)-3-(5-isobutyryl-2-methyl-4,5,6,7-tetrahydro-1H-imidazo[4,5-c]pyridin-1-yl)-8-azabicyclo[3.2.1]oct-8-yl]propyl}-acetamide (PF-232798). *J. Med. Chem.* **2011**, *54*, 67–77, DOI: 10.1021/jm100978n.

(110) Price, D. A.; Armour, D.; de Groot, M.; Leishman, D.; Napier, C.; Perros, M.; Stammen, B. L.; Wood, A. Overcoming HERG affinity in the discovery of the CCR5 antagonist maraviroc. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4633–4637, DOI: 10.1016/j.bmcl.2006.06.012.

(111) Ritchie, T. J.; Macdonald, S. J. F. Heterocyclic replacements for benzene: Maximising ADME benefits by considering individual ring isomers. *Eur. J. Med. Chem.* **2016**, *124*, 1057–1068, DOI: 10.1016/j.ejmech.2016.10.029.

(112) Johansson, A. et al. Discovery of (3-(4-(2-Oxa-6-azaspiro[3.3]heptan-6-ylmethyl)phenoxy)azetidin-1-yl)(5-(4-methoxyphenyl)-1,3,4-oxadiazol-2-yl)methanone (AZD1979), a Melanin Concentrating Hormone Receptor 1 (MCHr1) Antagonist with Favorable Physicochemical Properties. *J. Med. Chem.* **2016**, *59*, 2497–2511, DOI: 10.1021/acs.jmedchem.5b01654.

(113) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463–469, DOI: 10.1038/nature04710.

(114) Goldberg, K.; Groombridge, S.; Hudson, J.; Leach, A. G.; MacFaul, P. A.; Pickup, A.; Poultney, R.; Scott, J. S.; Svensson, P. H.; Sweeney, J. Oxadiazole isomers: all bioisosteres are not created equal. *MedChemComm* **2012**, *3*, 600, DOI: 10.1039/C2MD20054F.

(115) Boström, J.; Hogner, A.; Llinàs, A.; Wellner, E.; Plowright, A. T. Oxadiazoles in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 1817–1830, DOI: 10.1021/jm2013248.

(116) He, Q.; Yuan, Q.; Shan, H.; Wu, C.; Gu, Y.; Wu, K.; Hu, W.; Zhang, Y.; He, X.; Xu, H. E.; Zhao, L.-H. Mechanisms of ligand recognition and activation of melanin-concentrating hormone receptors. *Cell Discov.* **2024**, *10*, 48, DOI: 10.1038/s41421-024-00679-8.

(117) Rankovic, Z. CNS drug design: balancing physicochemical properties for optimal brain exposure. *J. Med. Chem.* **2015**, *58*, 2584–2608, DOI: 10.1021/jm501535r.

(118) Cosgrove, B.; Down, K.; Bertrand, S.; Tomkinson, N. C. O.; Barker, M. D. Investigating the effects of the core nitrogen atom configuration on the thermodynamic solubility of 6,5-bicyclic heterocycles. *Bioorg. Med. Chem. Lett.* **2021**, *33*, 127752, DOI: 10.1016/j.bmcl.2020.127752.

(119) Smith, D. A.; Allerton, C.; Kalgutkar, A. S.; van de Waterbee, H.; Walker, D. K., *Pharmacokinetics and Metabolism in Drug Design*, 3rd ed.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag: Weinheim, Germany, 2012.

(120) Zhang, Z.; Tang, W. Drug metabolism in drug discovery and development. *Acta Pharm. Sin. B.* **2018**, *8*, 721–732, DOI: 10.1016/j.apsb.2018.04.003.

(121) Iyer, K. R.; Sinz, M. W. Characterization of Phase I and Phase II hepatic drug metabolism activities in a panel of human liver preparations. *Chem. Biol. Interact.* **1999**, *118*, 151–169, DOI: 10.1016/S0009-2797(99)00007-1.

(122) Ingelman-Sundberg, M. Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. *Naunyn. Schmiedebergs. Arch. Pharmacol.* **2004**, *369*, 89–104, DOI: 10.1007/s00210-003-0819-z.

(123) Jancova, P.; Anzenbacher, P.; Anzenbacherova, E. Phase II drug metabolizing enzymes. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech. Repub.* **2010**, *154*, 103–116, DOI: 10.5507/bp.2010.017.

(124) Kosoglou, T.; Statkevich, P.; Johnson-Levonas, A. O.; Paolini, J. F.; Bergman, A. J.; Alton, K. B. Ezetimibe: a review of its metabolism, pharmacokinetics and drug interactions. *Clin. Pharmacokinet.* **2005**, *44*, 467–494, DOI: 10.2165/00003088-200544050-00002.

(125) Ogilvie, B. W.; Zhang, D.; Li, W.; Rodrigues, A. D.; Gipson, A. E.; Holsapple, J.; Toren, P.; Parkinson, A. Glucuronidation converts gemfibrozil to a potent, metabolism-dependent inhibitor of CYP2C8: implications for drug-drug interactions. *Drug Metab. Dispos.* **2006**, *34*, 191–197, DOI: 10.1124/dmd.105.007633.

(126) Guengerich, F. P.; Johnson, W. W.; Shimada, T; Ueng, Y. F.; Yamazaki, H; Langouët, S Activation and detoxication of aflatoxin B1. *Mutat. Res.* **1998**, *402*, 121–128, DOI: 10.1016/s0027-5107(97)00289-3.

(127) Mulders, T. M.; Venizelos, V; Schoemaker, R; Cohen, A. F.; Breimer, D. D.; Mulder, G. J. Characterization of glutathione conjugation in humans: stereoselectivity in plasma elimination pharmacokinetics and urinary excretion of (R)- and (S)-2-bromoisovalerylurea in healthy volunteers. *Clin. Pharmacol. Ther.* **1993**, *53*, 49–58, DOI: 10.1038/clpt.1993.8.

(128) Raghavan, N.; Frost, C. E.; Yu, Z.; He, K.; Zhang, H.; Humphreys, W. G.; Pinto, D.; Chen, S.; Bonacorsi, S.; Wong, P. C.; Zhang, D. Apixaban metabolism and phar-

macokinetics after oral administration to humans. *Drug Metab. Dispos.* **2009**, *37*, 74–81, DOI: 10.1124/dmd.108.023143.

(129) Anderson, R. J.; Kudlacek, P. E.; Clemens, D. L. Sulfation of minoxidil by multiple human cytosolic sulfotransferases. *Chem. Biol. Interact.* **1998**, *109*, 53–67, DOI: 10.1016/s0009-2797(97)00120-8.

(130) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Discov.* **2015**, *14*, 387–404, DOI: 10.1038/nrd4581.

(131) Park, B. K. et al. Managing the challenge of chemically reactive metabolites in drug development. *Nat. Rev. Drug Discov.* **2011**, *10*, 292–306, DOI: 10.1038/nrd3408.

(132) Smith, D.; Schmid, E.; Jones, B. Do drug metabolism and pharmacokinetic departments make any contribution to drug discovery? *Clin. Pharmacokinet.* **2002**, *41*, 1005–1019, DOI: 10.2165/00003088-200241130-00001.

(133) Cohen, J. S. Risks of troglitazone apparent before approval in USA. *Diabetologia* **2006**, *49*, 1454–1455, DOI: 10.1007/s00125-006-0245-0.

(134) Pfizer Pfizer Annual Report 2001, https://people.stern.nyu.edu/jbilders/Pdf/pfizer2001ar.pdf, [Accessed 27-03-2025].

(135) Kassahun, K; Pearson, P. G.; Tang, W; McIntosh, I; Leung, K; Elmore, C; Dean, D; Wang, R; Doss, G; Baillie, T. A. Studies on the metabolism of troglitazone to reactive intermediates in vitro and in vivo. Evidence for novel biotransformation pathways involving quinone methide formation and thiazolidinedione ring scission. *Chem. Res. Toxicol.* **2001**, *14*, 62–70, DOI: 10.1021/tx000180q.

(136) Bolton, J. L.; Trush, M. A.; Penning, T. M.; Dryhurst, G; Monks, T. J. Role of quinones in toxicology. *Chem. Res. Toxicol.* **2000**, *13*, 135–160, DOI: 10.1021/tx9902082.

(137) Masubuchi, Y. Metabolic and non-metabolic factors determining troglitazone hepatotoxicity: a review. *Drug Metab. Pharmacokinet.* **2006**, *21*, 347–356, DOI: 10.2133/dmpk.21.347.

(138) Arrowsmith, C. H. et al. The promise and peril of chemical probes. *Nat. Chem. Biol.* **2015**, *11*, 536–541, DOI: 10.1038/nchembio.1867.

(139) Smith, D. A.; Beaumont, K.; Maurer, T. S.; Di, L. Clearance in drug design. *J. Med. Chem.* **2019**, *62*, 2245–2255, DOI: 10.1021/acs.jmedchem.8b01263.

(140) Trunzer, M.; Faller, B.; Zimmerlin, A. Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. *J. Med. Chem.* **2009**, *52*, 329–335, DOI: 10.1021/jm8008663.

(141) Garcia-Perez, I.; Posma, J. M.; Serrano-Contreras, J. I.; Boulangé, C. L.; Chan, Q.; Frost, G.; Stamler, J.; Elliott, P.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. Identifying unknown metabolites using NMR-based metabolic profiling techniques. *Nat. Protoc.* **2020**, *15*, 2538–2567, DOI: 10.1038/s41596-020-0343-3.

(142) Thompson, T. N. Optimization of metabolic stability as a goal of modern drug design. *Med. Res. Rev.* **2001**, *21*, 412–449, DOI: 10.1002/med.1017.

(143) Lin, J. H. Role of pharmacokinetics in the discovery and development of indinavir. *Adv. Drug Deliv. Rev.* **1999**, *39*, 33–49, DOI: 10.1016/s0169-409x(99)00018-6.

(144) St Jean Jr, D. J.; Fotsch, C. Mitigating heterocycle metabolism in drug discovery. *J. Med. Chem.* **2012**, *55*, 6002–6020, DOI: 10.1021/jm300343m.

(145) Riether, D. et al. 1,4-Diazepane compounds as potent and selective CB2 agonists: optimization of metabolic stability. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 2011–2016, DOI: 10.1016/j.bmcl.2011.02.017.

(146) Ishida, H.; Isami, S.; Matsumura, T.; Umehara, H.; Yamashita, Y.; Kajita, J.; Fuse, E.; Kiyoi, H.; Naoe, T.; Akinaga, S.; Shiotsu, Y.; Arai, H. Novel and orally active

5-(1,3,4-oxadiazol-2-yl)pyrimidine derivatives as selective FLT3 inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5472–5477, DOI: 10.1016/j.bmcl.2008.09.031.

(147) Wishka, D. G. et al. Discovery of N-[(3R)-1-azabicyclo[2.2.2]oct-3-yl]furo[2,3-c]pyridine-5-carboxamide, an agonist of the alpha7 nicotinic acetylcholine receptor, for the potential treatment of cognitive deficits in schizophrenia: synthesis and structure–activity relationship. *J. Med. Chem.* **2006**, *49*, 4425–4436, DOI: 10.1021/jm0602413.

(148) Kalgutkar, A. S.; Gardner, I.; Obach, R. S.; Shaffer, C. L.; Callegari, E.; Henne, K. R.; Mutlib, A. E.; Dalvie, D. K.; Lee, J. S.; Nakai, Y.; O'Donnell, J. P.; Boer, J.; Harriman, S. P. A comprehensive listing of bioactivation pathways of organic functional groups. *Curr. Drug Metab.* **2005**, *6*, 161–225, DOI: 10.2174/1389200054021799.

(149) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52*, 6752–6756, DOI: 10.1021/jm901241e.

(150) Churcher, I.; Newbold, S.; Murray, C. W. Return to flatland. *Nat. Rev. Chem.* **2025**, *9*, 140–141, DOI: 10.1038/s41570-025-00688-5.

(151) Buskes, M. J.; Blanco, M.-J. Impact of cross-coupling reactions in drug discovery and development. *Molecules* **2020**, *25*, 3493, DOI: 10.3390/molecules25153493.

(152) *Bioisosteres in Medicinal Chemistry*; Brown, N., Ed.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag: Weinheim, Germany, 2012.

(153) Friedman, H. L. Influence of isosteric replacements upon biological activity. *NAS-NRS Publication* **1951**, *206*, 295–358.

(154) Patani, G. A.; LaVoie, E. J. Bioisosterism: a rational approach in drug design. *Chemi. Rev.* **1996**, *96*, 3147–3176, DOI: 10.1021/cr950066q.

(155) Lipinski, C. A. Bioisosterism in drug design, In *Annual Reports in Medicinal Chemistry*, ed. by Bailey, D. M., Elsevier: 1986, pp 283–291, DOI: 10.1016/S0065-7743(08)61137-9.

(156) Salmon, J. A.; Garland, L. G.; Hoyle, B. D.; Costerton, J. W.; Seiler, N.; Raeburn, D.; Karlsson, J.-A.; Polak, A; Hartman, P.; Rohmer, M., et al. Isosterism and bioisosterism in drug design, In *Progress in Drug Research/Fortschritte der Arzneimittelforschung/Progrès des recherches pharmaceutiques*, Springer: 1991, pp 287–371, DOI: 10.1007/978-3-0348-7139-6_7.

(157) Langmuir, I. Isomorphism, Isosterism and Covalence. *Journal of the American Chemical Society* **1919**, *41*, 1543–1559, DOI: 10.1021/ja02231a009.

(158) Burger, A. Isosterism and bioisosterism in drug design, In *Progress in Drug Research / Fortschritte der Arzneimittelforschung / Progrès des recherches pharmaceutiques*, Birkhäuser Basel: Basel, 1991, pp 287–371.

(159) Grimm, H. G. Zur Systematik der chemischen Verbindungen vom Standpunkt der Atomforschung, zugleich über einige Aufgaben der Experimentalchemie. *Sci. Nat.* **1929**, *17*, 557–564, DOI: 10.1007/BF01505929.

(160) Thornber, C. W. Isosterism and molecular modification in drug design. *Chem. Soc. Rev.* **1979**, *8*, 563–580, DOI: 10.1039/CS9790800563.

(161) Wermuth, C. G., *The Practice of Medicinal Chemistry*, 3rd ed.; Wermuth, C. G., Ed.; Academic Press: 2011.

(162) Herr, R. J. 5-Substituted-1H-tetrazoles as carboxylic acid isosteres: medicinal chemistry and synthetic methods. *Bioorg. Med. Chem.* **2002**, *10*, 3379–3393, DOI: 10.1016/s0968-0896(02)00239-0.

(163) Chen, X.; Wang, W. Chapter 32. The use of bioisosteric groups in lead optimization, In *Annual Reports in Medicinal Chemistry*, Elsevier: 2003, pp 333–346, DOI: 10.1016/S0065-7743(03)38033-9.

(164) Wermuth, C. G. Similarity in drugs: reflections on analogue design. *Drug Discov. Today* **2006**, *11*, 348–354, DOI: 10.1016/j.drudis.2006.02.006.

(165) Callis, T. B.; Garrett, T. R.; Montgomery, A. P.; Danon, J. J.; Kassiou, M. Recent scaffold hopping applications in central nervous system drug discovery. *J. Med. Chem.* **2022**, *65*, 13483–13504, DOI: 10.1021/acs.jmedchem.2c00969.

(166) Gu, H.; Zhang, S. Advances in kinetic isotope effect measurement techniques for enzyme mechanism study. *Molecules* **2013**, *18*, 9278–9292, DOI: 10.3390/molecules18089278.

(167) Willistein, M.; Bechtel, D. F.; Müller, C. S.; Demmer, U.; Heimann, L.; Kayastha, K.; Schünemann, V.; Pierik, A. J.; Ullmann, G. M.; Ermler, U.; Boll, M. Low potential enzymatic hydride transfer via highly cooperative and inversely functionalized flavin cofactors. *Nat. Commun.* **2019**, *10*, 2074, DOI: 10.1038/s41467-019-10078-3.

(168) Meanwell, N. A. The influence of bioisosteres in drug design: tactical applications to address developability problems, In *Tactics in Contemporary Drug Design*, ed. by Meanwell, N. A., Springer Berlin Heidelberg: Berlin, Heidelberg, 2013, pp 283–381, DOI: 10.1007/7355_2013_29.

(169) Truong, T. M.; Pathak, G. N.; Singal, A.; Taranto, V.; Rao, B. K. Deucravacitinib: The first FDA-approved oral TYK2 inhibitor for moderate to severe plaque psoriasis. *Ann. Pharmacother.* **2024**, *58*, 416–427, DOI: 10.1177/10600280231153863.

(170) Moslin, R. et al. Identification of N-methyl nicotinamide and N-methyl pyridazine-3-carboxamide pseudokinase domain ligands as highly selective allosteric inhibitors

of tyrosine kinase 2 (TYK2). *J. Med. Chem.* **2019**, *62*, 8953–8972, DOI: 10.1021/acs.jmedchem.9b00443.

(171) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Devel.* **2007**, *10*, 316–324.

(172) Phillips, G; Davey, D. D.; Eagen, K. A.; Koovakkat, S. K.; Liang, A; Ng, H. P.; Pinkerton, M; Trinh, L; Whitlow, M; Beatty, A. M.; Morrissey, M. M. Design, synthesis, and activity of 2,6-diphenoxypyridine-derived factor Xa inhibitors. *J. Med. Chem.* **1999**, *42*, 1749–1756, DOI: 10.1021/jm980667k.

(173) Penning, T. D. et al. Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benze nesulfonamide (SC-58635, celecoxib). *J. Med. Chem.* **1997**, *40*, 1347–1365, DOI: 10.1021/jm960803q.

(174) Phillips, G. B. et al. Discovery of N-[2-[5-[Amino(imino)methyl]-2-hydroxyphenoxy]-3, 5-difluoro-6-[3-(4, 5-dihydro-1-methyl-1H-imidazol-2-yl)phenoxy]pyridin-4-yl]-N-methylgl y cine (ZK-807834): a potent, selective, and orally active inhibitor of the blood coagulation enzyme factor Xa. *J. Med. Chem.* **1998**, *41*, 3557–3562, DOI: 10.1021/jm980280h.

(175) Adler, M; Davey, D. D.; Phillips, G. B.; Kim, S. H.; Jancarik, J; Rumennik, G; Light, D. R.; Whitlow, M Preparation, characterization, and the crystal structure of the inhibitor ZK-807834 (CI-1031) complexed with factor Xa. *Biochemistry* **2000**, *39*, 12534–12542, DOI: 10.1021/bi001477q.

(176) Gadiya, Y.; Gribbon, P.; Hofmann-Apitius, M.; Zaliani, A. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery per-

spective. *Artif. Intell. Life Sci.* **2023**, *3*, 100069, DOI: 10.1016/j.ailsci.2023.100069.

(177) Aronson, J. K.; Green, A. R. Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists. *Br. J. Clin. Pharmacol.* **2020**, *86*, 2114–2122, DOI: 10.1111/bcp.14327.

(178) Khanna, I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov. Today* **2012**, *17*, 1088–1102, DOI: 10.1016/j.drudis.2012.05.007.

(179) Stewart, K. D.; Shanley, J.; Ahmed, K. B. A.; Bowen, J. P. The Drug Guru Project, In *Bioisosteres in Medicinal Chemistry*, ed. by Brown, N., Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2012, pp 183–198, DOI: 10.1002/9783527654307.ch11.

(180) Haning, H.; Niewöhner, U.; Schenke, T.; Es-Sayed, M.; Schmidt, G.; Lampe, T.; Bischoff, E. Imidazo[5,1- f] [1,2,4] triazin-4(3 H)-ones, a new class of potent PDE 5 inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 865–868, DOI: 10.1016/s0960-894x(02)00030-6.

(181) Ghofrani, H. A.; Osterloh, I. H.; Grimminger, F. Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nat. Rev. Drug Discov.* **2006**, *5*, 689–702, DOI: 10.1038/nrd2030.

(182) Laties, A.; Zrenner, E. Viagra (sildenafil citrate) and ophthalmology. *Prog. Retin. Eye Res.* **2002**, *21*, 485–506, DOI: 10.1016/s1350-9462(02)00013-7.

(183) Keating, G. M.; Scott, L. J. Vardenafil: a review of its use in erectile dysfunction. *Drugs* **2003**, *63*, 2673–2703, DOI: 10.2165/00003495-200363230-00010.

(184) Wang, H.; Ye, M.; Robinson, H.; Francis, S. H.; Ke, H. Conformational variations of both phosphodiesterase-5 and inhibitors provide the structural basis for the physio-

logical effects of vardenafil and sildenafil. *Mol. Pharmacol.* **2008**, *73*, 104–110, DOI: 10.1124/mol.107.040212.

(185) Rubio-Aurioles, E.; Porst, H.; Eardley, I.; Goldstein, I. Comparing vardenafil and sildenafil in the treatment of men with erectile dysfunction and risk factors for cardiovascular disease: a randomized, double-blind, pooled crossover study. *J. Sex. Med.* **2006**, *3*, 1037–1049, DOI: 10.1111/j.1743-6109.2006.00310.x.

(186) Saenz de Tejada, I; Angulo, J; Cuevas, P; Fernández, A; Moncada, I; Allona, A; Lledó, E; Körschen, H. G.; Niewöhner, U; Haning, H; Pages, E; Bischoff, E The phosphodiesterase inhibitory selectivity and the in vitro and in vivo potency of the new PDE5 inhibitor vardenafil. *Int. J. Impot. Res.* **2001**, *13*, 282–290, DOI: 10.1038/sj.ijir.3900726.

(187) Corbin, J. D.; Beasley, A.; Blount, M. A.; Francis, S. H. Vardenafil: structural basis for higher potency over sildenafil in inhibiting cGMP-specific phosphodiesterase-5 (PDE5). *Neurochem. Int.* **2004**, *45*, 859–863, DOI: 10.1016/j.neuint.2004.03.016.

(188) Mehrotra, N; Gupta, M; Kovar, A; Meibohm, B The role of pharmacokinetics and pharmacodynamics in phosphodiesterase-5 inhibitor therapy. *Int. J. Impot. Res.* **2007**, *19*, 253–264, DOI: 10.1038/sj.ijir.3901522.

(189) Ballatore, C.; Huryn, D. M.; Smith 3rd, A. B. Carboxylic acid (bio)isosteres in drug design. *ChemMedChem* **2013**, *8*, 385–395, DOI: 10.1002/cmdc.201200585.

(190) Noda, K; Saad, Y; Kinoshita, A; Boyle, T. P.; Graham, R. M.; Husain, A; Karnik, S. S. Tetrazole and carboxylate groups of angiotensin receptor antagonists bind to the same subsite by different mechanisms. *J. Biol. Chem.* **1995**, *270*, 2284–2289, DOI: 10.1074/jbc.270.5.2284.

(191) Lassalas, P.; Gay, B.; Lasfargeas, C.; James, M. J.; Tran, V.; Vijayendran, K. G.; Brunden, K. R.; Kozlowski, M. C.; Thomas, C. J.; Smith 3rd, A. B.; Huryn, D. M.; Ballatore, C. Structure property relationships of carboxylic acid isosteres. *J. Med. Chem.* **2016**, *59*, 3183–3203, DOI: 10.1021/acs.jmedchem.5b01963.

(192) Hansch, C.; Leo, A.; Hoekman, D., *Exploring QSAR.: Hydrophobic, electronic, and steric constants*; ACS professional reference book, Vol. 2; American Chemical Society: 1995.

(193) Cuozzo, A.; Daina, A.; Perez, M. A. S.; Michielin, O.; Zoete, V. SwissBioisostere 2021: updated structural, bioactivity and physicochemical data delivered by a reshaped web interface. *Nucleic Acids Res.* **2021**, *50*, D1382–D1390, DOI: 10.1093/nar/gkab1047.

(194) Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **2024**, *52*, D1180–D1192, DOI: 10.1093/nar/gkad1004.

(195) Carini, D. J.; Duncia, J. V.; Aldrich, P. E.; Chiu, A. T.; Johnson, A. L.; Pierce, M. E.; Price, W. A.; Santella 3rd, J. B.; Wells, G. J.; Wexler, R. R. Nonpeptide angiotensin II receptor antagonists: the discovery of a series of N-(biphenylylmethyl)imidazoles as potent, orally active antihypertensives. *J. Med. Chem.* **1991**, *34*, 2525–2547, DOI: 10.1021/jm00112a031.

(196) Meanwell, N. A. Fluorine and fluorinated motifs in the design and application of bioisosteres for drug design. *J. Med. Chem.* **2018**, *61*, 5822–5880, DOI: 10.1021/acs.jmedchem.7b01788.

(197) Kumari, S.; Carmona, A. V.; Tiwari, A. K.; Trippier, P. C. Amide bond bioisosteres: Strategies, synthesis, and successes. *J. Med. Chem.* **2020**, *63*, 12290–12358, DOI: 10.1021/acs.jmedchem.0c00530.

(198) Subbaiah, M. A. M.; Meanwell, N. A. Bioisosteres of the phenyl ring: Recent strategic applications in lead optimization and drug design. *J. Med. Chem.* **2021**, *64*, 14046–14128, DOI: 10.1021/acs.jmedchem.1c01215.

(199) Meanwell, N. A. Applications of bioisosteres in the design of biologically active compounds. *J. Agric. Food Chem.* **2023**, *71*, 18087–18122, DOI: 10.1021/acs.jafc.3c00765.

(200) Ertl, P.; Altmann, E.; Racine, S. The most common linkers in bioactive molecules and their bioisosteric replacement network. *Bioorg. Med. Chem.* **2023**, *81*, 117194, DOI: 10.1016/j.bmc.2023.117194.

(201) Tsien, J.; Hu, C.; Merchant, R. R.; Qin, T. Three-dimensional saturated C(sp3)-rich bioisosteres for benzene. *Nat. Rev. Chem.* **2024**, *8*, 605–627, DOI: 10.1038/s41570-024-00623-0.

(202) Ertl, P. Database of 4 million medicinal chemistry-relevant ring systems. *J. Chem. Inf. Model.* **2024**, *64*, 1245–1250, DOI: 10.1021/acs.jcim.3c01812.

(203) Ujváry, I.; Hayward, J. Bioster: A Database of Bioisosteres and Bioanalogues, In *Bioisosteres in Medicinal Chemistry*, ed. by Brown, N., Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2012, pp 53–74, DOI: 10.1002/9783527654307.ch4.

(204) wwPDB consortium Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47*, D520–D528, DOI: 10.1093/nar/gky949.

(205) PDB statistics, https://www.rcsb.org/stats, [Accessed 30-03-2025].

(206) Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttmann, C. Identification of target-specific bioisosteric fragments from ligand-protein crystallographic data. *J. Comput. Aided Mol. Des.* **2006**, *20*, 385–394, DOI: 10.1007/s10822-006-9072-0.

(207) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615–633, DOI: 10.1016/0898-5529(90)90162-2.

(208) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031–2043, DOI: 10.1021/ci3000776.

(209) Desaphy, J.; Rognan, D. sc-PDB-Frag: a database of protein-ligand interaction patterns for Bioisosteric replacements. *J. Chem. Inf. Model.* **2014**, *54*, 1908–1918, DOI: 10.1021/ci500282c.

(210) Khashan, R. FragVLib a free database mining software for generating "Fragment-based Virtual Library" using pocket similarity search of ligand-receptor complexes. *J. Cheminform.* **2012**, *4*, 18, DOI: 10.1186/1758-2946-4-18.

(211) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **2008**, *3*, 435–444, DOI: 10.1002/cmdc.200700139.

(212) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522, DOI: 10.1021/ci970429i.

(213) Konc, J.; Janezic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168, DOI: 10.1093/bioinformatics/btq100.

(214) Lešnik, S.; Škrlj, B.; Eržen, N.; Bren, U.; Gobec, S.; Konc, J.; Janežič, D. BoBER: web interface to the base of bioisosterically exchangeable replacements. *J. Cheminform.* **2017**, *9*, 62, DOI: 10.1186/s13321-017-0251-x.

(215) Zhang, T.; Sun, S.; Wang, R.; Li, T.; Gan, B.; Zhang, Y. BioisoIdentifier: an online free tool to investigate local structural replacements from PDB. *J. Cheminform.* **2024**, *16*, 7, DOI: 10.1186/s13321-024-00801-8.

(216) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: molecular overlay based on shape and electrostatic potential. *J. Chem. Inf. Model.* **2009**, *49*, 492–502, DOI: 10.1021/ci800315d.

(217) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620, DOI: 10.1093/nar/gkv352.

(218) ChEMBL 35, https://www.ebi.ac.uk/chembl/, [Accessed 31-03-2025].

(219) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* **2012**, *41*, D1137–D1143, DOI: 10.1093/nar/gks1059.

(220) Alessandro, C.; Antoine, D.; Marta A S, P.; Olivier, M.; Vincent, Z. SwissBioisostere 2021: updated structural, bioactivity and physicochemical data delivered by a reshaped web interface. *Nucleic Acids Res.* **2022**, *50*, D1382–D1390, DOI: 10.1093/nar/gkab1047.

(221) Yang, Z.; Shi, S.; Fu, L.; Lu, A.; Hou, T.; Cao, D. Matched molecular pair analysis in drug discovery: Methods and recent applications. *J. Med. Chem.* **2023**, *66*, 4361–4377, DOI: 10.1021/acs.jmedchem.2c01787.

(222) Pastor, J. et al. Hit to lead evaluation of 1,2,3-triazolo[4,5-b]pyridines as PIM kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 1591–1597, DOI: 10.1016/j.bmcl.2011.12.130.

(223) Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348, DOI: 10.1021/ci900450m.

(224) Oliveira, T.; Silva, M.; Maia, E.; Silva, A.; Taranto, A. Virtual screening algorithms in drug discovery: A review focused on machine and deep learning methods. *Drugs Drug Candidates* **2023**, *2*, 311–334, DOI: 10.3390/ddc2020017.

(225) Gentile, F.; Yaacoub, J. C.; Gleave, J.; Fernandez, M.; Ton, A.-T.; Ban, F.; Stern, A.; Cherkasov, A. Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **2022**, *17*, 672–697, DOI: 10.1038/s41596-021-00659-2.

(226) Gangwal, A.; Lavecchia, A. Unleashing the power of generative AI in drug discovery. *Drug Discov. Today* **2024**, *29*, 103992, DOI: 10.1016/j.drudis.2024.103992.

(227) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* **2012**, *14*, 133–141, DOI: 10.1208/s12248-012-9322-0.

(228) Vázquez, J.; López, M.; Gibert, E.; Herrero, E.; Luque, F. J. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules* **2020**, *25*, 4723, DOI: 10.3390/molecules25204723.

(229) Maia, E. H. B.; Assis, L. C.; de Oliveira, T. A.; da Silva, A. M.; Taranto, A. G. Structure-based virtual screening: From classical to artificial intelligence. *Front. Chem.* **2020**, *8*, 343, DOI: 10.3389/fchem.2020.00343.

(230) Zhu, H.; Zhang, Y.; Li, W.; Huang, N. A comprehensive survey of prospective structure-based virtual screening for early drug discovery in the past fifteen years. *Int. J. Mol. Sci.* **2022**, *23*, 15961, DOI: 10.3390/ijms232415961.

(231) Zhao, H. The science and art of structure-based virtual screening. *ACS Med. Chem. Lett.* **2024**, *15*, 436–440, DOI: 10.1021/acsmedchemlett.4c00093.

(232) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63, DOI: 10.1016/j.ymeth.2014.08.005.

(233) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771–784, DOI: 10.1021/ci100062n.

(234) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079–2093, DOI: 10.1021/ci100263p.

(235) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280, DOI: 10.1021/ci010132r.

(236) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated platform of small molecules and biological activities, In *Annual Reports in Computational Chemistry*, ed. by Wheeler, R. A.; Spellmeyer, D. C., Elsevier: 2008, pp 217–241, DOI: 10.1016/S1574-1400(08)00012-1.

(237) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754, DOI: 10.1021/ci100050t.

(238) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151–1157, DOI: 10.1021/ci030285+.

(239) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol. Inform.* **2010**, *29*, 366–385, DOI: 10.1002/minf.201000019.

(240) Lauri, G; Bartlett, P. A. CAVEAT: a program to facilitate the design of organic molecules. *J. Comput. Aided Mol. Des.* **1994**, *8*, 51–66, DOI: 10.1007/BF00124349.

(241) Rush 3rd, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495, DOI: 10.1021/jm040163o.

(242) Tu, M.; Rai, B. K.; Mathiowetz, A. M.; Didiuk, M.; Pfefferkorn, J. A.; Guzman-Perez, A.; Benbow, J.; Guimarães, C. R. W.; Mente, S.; Hayward, M. M.; Liras, S. Exploring aromatic chemical space with NEAT: novel and electronically equivalent aromatic template. *J. Chem. Inf. Model.* **2012**, *52*, 1114–1123, DOI: 10.1021/ci300031s.

(243) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875, DOI: 10.1021/ci300415d.

(244) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801, DOI: 10.1021/jm0608356.

(245) Taminau, J.; Thijs, G.; De Winter, H. Pharao: pharmacophore alignment and optimization. *J. Mol. Graph. Model.* **2008**, *27*, 161–169, DOI: 10.1016/j.jmgm.2008.04.003.

(246) Roy, A.; Skolnick, J. LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. *Bioinformatics* **2015**, *31*, 539–544, DOI: 10.1093/bioinformatics/btu692.

(247) Hu, J.; Liu, Z.; Yu, D.-J.; Zhang, Y. LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics* **2018**, *34*, 2209–2218, DOI: 10.1093/bioinformatics/bty081.

(248) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723, DOI: 10.1002/jcc.20681.

(249) Ballester, P. J.; Finn, P. W.; Richards, W. G. Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J. Mol. Graph. Model.* **2009**, *27*, 836–845, DOI: 10.1016/j.jmgm.2009.01.001.

(250) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided Mol. Des.* **2010**, *24*, 789–801, DOI: 10.1007/s10822-010-9374-0.

(251) Takács, G.; Havasi, D.; Sándor, M.; Dohánics, Z.; Balogh, G. T.; Kiss, R. DIY virtual chemical libraries - novel starting points for drug discovery. *ACS Med. Chem. Lett.* **2023**, *14*, 1188–1197, DOI: 10.1021/acsmedchemlett.3c00146.

(252) Kennedy, J. P.; Williams, L.; Bridges, T. M.; Daniels, R. N.; Weaver, D.; Lindsley, C. W. Application of combinatorial chemistry science on modern drug discovery. *J. Comb. Chem.* **2008**, *10*, 345–354, DOI: 10.1021/cc700187t.

(253) MCule Ultimate, https://ultimate.mcule.com/, [Accessed 31-03-2025].

(254) WuXi AppTec Galaxi, https://www.wuxiapptec.com/news/wuxi-news/3486, [Accessed 31-03-2025].

(255) Enamine REAL, https://enamine.net/compound-collections/real-compounds/real-database, [Accessed 31-03-2025].

(256) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic rings of the future. *J. Med. Chem.* **2009**, *52*, PMID: 19348472, 2952–2963, DOI: 10.1021/jm801513z.

(257) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8, DOI: 10.1186/1758-2946-1-8.

(258) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

(259) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, DOI: 10.1002/wcms.1603.

(260) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. AutodE: automated calculation of reaction energy profiles- application to organic and organometallic reactions. *Angew. Chem. Int. Ed.* **2021**, *60*, 4266–4274, DOI: 10.1002/anie.202011941.

(261) Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ($Z = 1$–86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009, DOI: 10.1021/acs.jctc.7b00118.

(262) Schlegel, H. B. Geometry optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 790–809, DOI: 10.1002/wcms.34.

(263) Weiner, P. K.; Langridge, R; Blaney, J. M.; Schaefer, R; Kollman, P. A. Electrostatic potential molecular surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 3754–3758, DOI: 10.1073/pnas.79.12.3754.

(264) Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with aromatic rings in chemical and biological recognition. *Angew. Chem. Int. Ed.* **2003**, *42*, 1210–1250, DOI: 10.1002/anie.200390319.

(265) Cho, M.; Sylvetsky, N.; Eshafi, S.; Santra, G.; Efremenko, I.; Martin, J. M. L. The atomic partial charges arboretum: trying to see the forest for the trees. *Chemphyschem* **2020**, *21*, 688–696, DOI: 10.1002/cphc.202000040.

(266) Liu, S. In Liu, S., Ed.; Blackwell Verlag: Berlin, Germany, 2024; Chapter Partial Charges.

(267) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094, DOI: 10.1021/cr9904009.

(268) Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161–2200, DOI: 10.1021/cr960149m.

(269) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **2020**, *12*, 56, DOI: 10.1186/s13321-020-00460-5.

(270) Grant, J. A.; Pickup, B. T. A Gaussian description of molecular shape. *J. Phys. Chem.* **1995**, *99*, 3503–3510, DOI: 10.1021/j100011a016.

(271) Krissinel, E. B.; Henrick, K. Common subgraph isomorphism detection by backtracking search. *Softw. Pract. Exp.* **2004**, *34*, 591–607, DOI: 10.1002/spe.588.

(272) Kavan, L.; Collins, S.; O'Sullivan, C.; Zara, J. *Dual quaternions for rigid transformation blending*; tech. rep. TCD-CS-2006-46; Trinity College Dublin, 2006.

(273) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188–191, DOI: 10.1021/ci00007a002.

(274) Boby, M. L. et al. Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science* **2023**, *382*, eabo7201, DOI: 10.1126/science.abo7201.

(275) World Health Organization 2023 data.who.int, WHO Coronavirus (COVID-19) dashboard, https://data.who.int/dashboards/covid19, [Accessed 19-02-2025].

(276) Ullrich, S.; Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* **2020**, *30*, 127377, DOI: 10.1016/j.bmcl.2020.127377.

(277) Malone, B.; Urakova, N.; Snijder, E. J.; Campbell, E. A. Structures and functions of coronavirus replication-transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 21–39, DOI: 10.1038/s41580-021-00432-z.

(278) Jin, Z. et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289–293, DOI: 10.1038/s41586-020-2223-y.

(279) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375, DOI: 10.1021/ci0500177.

(280) Macip, G.; Garcia-Segura, P.; Mestres-Truyol, J.; Saldivar-Espinoza, B.; Ojeda-Montes, M. J.; Gimeno, A.; Cereto-Massagué, A.; Garcia-Vallvé, S.; Pujadas, G. Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. *Med. Res. Rev.* **2022**, *42*, 744–769, DOI: 10.1002/med.21862.

(281) Bzówka, M.; Mitusińska, K.; Raczyńska, A.; Samol, A.; Tuszyński, J. A.; Góra, A. Structural and evolutionary analysis indicate that the SARS-CoV-2 Mpro is a challenging target for small-molecule inhibitor design. *Int. J. Mol. Sci.* **2020**, *21*, 3099, DOI: 10.3390/ijms21093099.

(282) Zev, S.; Raz, K.; Schwartz, R.; Tarabeh, R.; Gupta, P. K.; Major, D. T. Benchmarking the ability of common docking programs to correctly reproduce and score binding modes in SARS-CoV-2 protease Mpro. *J. Chem. Inf. Model.* **2021**, *61*, 2957–2966, DOI: 10.1021/acs.jcim.1c00263.

(283) Hu, Y.; Stumpfe, D.; Bajorath, J. Computational exploration of molecular scaffolds in medicinal chemistry. *J. Med. Chem.* **2016**, *59*, PMID: 26840095, 4062–4076, DOI: 10.1021/acs.jmedchem.5b01746.

(284) Ertl, P. Database of bioactive ring systems with calculated properties and its use in bioisosteric design and scaffold hopping. *Bioorg. Med. Chem.* **2012**, *20*, Chemoinformatics in Drug Discovery, 5436–5442, DOI: 10.1016/j.bmc.2012.02.058.

(285) Gavaghan, D. Problems with the current approach to the dissemination of computational science research and its implications for research integrity. *Bull. Math. Biol.* **2018**, *80*, 3088–3094, DOI: 10.1007/s11538-018-0499-y.

(286) Osborne, J. M.; Bernabeu, M. O.; Bruna, M.; Calderhead, B.; Cooper, J.; Dalchau, N.; Dunn, S.-J.; Fletcher, A. G.; Freeman, R.; Groen, D., et al. Ten simple rules for effective computational research. *PLOS Comput. Biol.* **2014**, *10*, e1003506, DOI: 10.1371/journal.pcbi.1003506.

(287) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465, DOI: 10.1002/jcc.21759.

(288) Bruns, R. F.; Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **2012**, *55*, PMID: 23061697, 9763–9772, DOI: 10.1021/jm301008n.

(289) Hammett, L. P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103, DOI: 10.1021/ja01280a022.

(290) Hall, R. J.; Murray, C. W.; Verdonk, M. L. The Fragment Network: A chemistry recommendation engine built using a graph database. *J. Med. Chem.* **2017**, *60*, 6440–6450, DOI: 10.1021/acs.jmedchem.7b00809.

(291) Nepali, K.; Lee, H.-Y.; Liou, J.-P. Nitro-group-containing drugs. *J. Med. Chem.* **2019**, *62*, 2851–2893, DOI: 10.1021/acs.jmedchem.8b00147.

(292) Yale, H. L. The trifluoromethyl group in medical chemistry. *J. Med. Chem.* **1959**, *1*, 121–133.

(293) Abula, A.; Xu, Z.; Zhu, Z.; Peng, C.; Chen, Z.; Zhu, W.; Aisa, H. A. Substitution effect of the trifluoromethyl group on the bioactivity in medicinal chemistry: Statistical analysis and energy calculations. *J. Chem. Inf. Model.* **2020**, *60*, 6242–6250, DOI: 10.1021/acs.jcim.0c00898.

(294) Shearer, J.; Castro, J. L.; Lawson, A. D. G.; MacCoss, M.; Taylor, R. D. Rings in clinical trials and drugs: Present and future. *J. Med. Chem.* **2022**, *65*, 8699–8712, DOI: 10.1021/acs.jmedchem.2c00473.

(295) Hunt, A.; Thomas, D., *The pragmatic programmer*; Addison Wesley: Boston, MA, 1999.

(296) Zelle, J. M., *Python programming*, 3rd ed.; Franklin, Beedle & Associates: Wilsonville, OR, 2016.

(297) Riniker, S.; Landrum, G. A. Better informed distance geometry: Using what we know to improve conformation generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574, DOI: 10.1021/acs.jcim.5b00654.

(298) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminform.* **2014**, *6*, 37, DOI: 10.1186/s13321-014-0037-3.

(299) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-practice DFT protocols for basic molecular computational chemistry. *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735, DOI: h10.1002/anie.202205735.

(300) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104, DOI: 10.1063/1.3382344.

(301) Neese, F. Software update: the ORCA program system, version 5.0. *WIRES Comput. Molec. Sci.* **2022**, *12*, e1606, DOI: 10.1002/wcms.1606.

(302) Kim, S.; Bolton, E. E.; Bryant, S. H. PubChem3D: conformer ensemble accuracy. *J. Cheminform.* **2013**, *5*, 1, DOI: 10.1186/1758-2946-5-1.

(303) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228, DOI: 10.1016/0040-4020(80)80168-2.

(304) North, S. C.; Jorgensen, K. R.; Pricetolstoy, J.; Wilson, A. K. Population analysis and the effects of Gaussian basis set quality and quantum mechanical approach: main group through heavy element species. *Front. Chem.* **2023**, *11*, 1152500, DOI: 10.3389/fchem.2023.1152500.

(305) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82, DOI: 10.1021/jm0603365.

(306) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1666, DOI: 10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K.

(307) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204, DOI: 10.1021/jm401411z.

(308) Bolcato, G.; Heid, E.; Boström, J. On the value of using 3D shape and electrostatic similarities in deep generative methods. *J. Chem. Inf. Model.* **2022**, *62*, 1388–1398, DOI: 10.1021/acs.jcim.1c01535.

(309) Kabsch, W A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **1976**, *32*, 922–923, DOI: 10.1107/S0567739476001873.

(310) Kabsch, W A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **1978**, *34*, 827–828, DOI: 10.1107/S0567739478001680.

(311) Drilon, A. et al. Repotrectinib in *ROS1* fusion-positive non-small-cell lung cancer. *N. Engl. J. Med.* **2024**, *390*, 118–131, DOI: 10.1056/NEJMoa2302299.

(312) Liu, Z.; Yu, P.; Dong, L.; Wang, W.; Duan, S.; Wang, B.; Gong, X.; Ye, L.; Wang, H.; Tian, J. Discovery of the next-generation pan-TRK kinase inhibitors for the treatment of cancer. *J. Med. Chem.* **2021**, *64*, PMID: 34253025, 10286–10296, DOI: 10.1021/acs.jmedchem.1c00712.

(313) Inc., L. O. Form 8-K, Securities and Exchange Commission EDGAR Database, 2023.

(314) Augtyro, https://www.bmspricinginformation.com/augtyro#, [Accessed 30-01-2025], 2024.

(315) Akiyama, H et al. Inflammation and Alzheimer's disease. *Neurobiol. Aging* **2000**, *21*, 383–421, DOI: 10.1016/s0197-4580(00)00124-x.

(316) Kinney, J. W.; Bemiller, S. M.; Murtishaw, A. S.; Leisgang, A. M.; Salazar, A. M.; Lamb, B. T. Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement. (N. Y.)* **2018**, *4*, 575–590, DOI: 10.1016/j.trci.2018.06.014.

(317) Leng, F.; Edison, P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nat. Rev. Neurol.* **2021**, *17*, 157–172, DOI: 10.1038/s41582-020-00435-y.

(318) Tansey, M. G.; Wallings, R. L.; Houser, M. C.; Herrick, M. K.; Keating, C. E.; Joers, V. Inflammation and immune dysfunction in Parkinson disease. *Nat. Rev. Immunol.* **2022**, *22*, 657–673, DOI: 10.1038/s41577-022-00684-6.

(319) Heneka, M. T. et al. Neuroinflammation in Alzheimer disease. *Nat. Rev. Immunol.* **2024**, *25*, 321–352, DOI: 10.1038/s41577-024-01104-7.

(320) Heneka, M. T.; McManus, R. M.; Latz, E. Inflammasome signalling in brain function and neurodegenerative disease. *Nat. Rev. Neurosci.* **2018**, *19*, 610–621, DOI: 10.1038/s41583-018-0055-7.

(321) Venegas, C. et al. Microglia-derived ASC specks cross-seed amyloid-β in Alzheimer's disease. *Nature* **2017**, *552*, 355–361, DOI: 10.1038/nature25158.

(322) Ising, C. et al. NLRP3 inflammasome activation drives tau pathology. *Nature* **2019**, *575*, 669–673, DOI: 10.1038/s41586-019-1769-z.

(323) Ising, C.; Heneka, M. T. Functional and structural damage of neurons by innate immune mechanisms during neurodegeneration. *Cell Death Dis.* **2018**, *9*, 120, DOI: 10.1038/s41419-017-0153-x.

(324) Yiannopoulou, K. G.; Anastasiou, A. I.; Zachariou, V.; Pelidou, S.-H. Reasons for failed trials of disease-modifying treatments for Alzheimer disease and their contribution in recent research. *Biomedicines* **2019**, *7*, 97, DOI: 10.3390/biomedicines7040097.

(325) Vande Walle, L.; Lamkanfi, M. Drugging the NLRP3 inflammasome: from signalling mechanisms to therapeutic targets. *Nat. Rev. Drug Discov.* **2024**, *23*, 43–66, DOI: 10.1038/s41573-023-00822-2.

(326) Swanson, K. V.; Deng, M.; Ting, J. P.-Y. The NLRP3 inflammasome: molecular activation and regulation to therapeutics. *Nat. Rev. Immunol.* **2019**, *19*, 477–489, DOI: 10.1038/s41577-019-0165-0.

(327) Yao, J.; Wang, Z.; Song, W.; Zhang, Y. Targeting NLRP3 inflammasome for neurodegenerative disorders. *Mol. Psychiatry* **2023**, *28*, 4512–4527, DOI: 10.1038/s41380-023-02239-0.

(328) Li, N.; Zhang, R.; Tang, M.; Zhao, M.; Jiang, X.; Cai, X.; Ye, N.; Su, K.; Peng, J.; Zhang, X.; Wu, W.; Ye, H. Recent progress and prospects of small molecules for NLRP3 inflammasome inhibition. *J. Med. Chem.* **2023**, *66*, 14447–14473, DOI: 10.1021/acs.jmedchem.3c01370.

(329) Nizami, S.; Millar, V.; Arunasalam, K.; Zarganes-Tzitzikas, T.; Brough, D.; Tresadern, G.; Brennan, P. E.; Davis, J. B.; Ebner, D.; Di Daniel, E. A phenotypic high-content, high-throughput screen identifies inhibitors of NLRP3 inflammasome activation. *Sci. Rep.* **2021**, *11*, 15319, DOI: 10.1038/s41598-021-94850-w.

(330) Dalke, A.; Hert, J.; Kramer, C. Mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *J. Chem. Inf. Model.* **2018**, *58*, 902–910, DOI: 10.1021/acs.jcim.8b00173.

(331) Coll, R. C. et al. A small-molecule inhibitor of the NLRP3 inflammasome for the treatment of inflammatory diseases. *Nat. Med.* **2015**, *21*, 248–255, DOI: 10.1038/nm.3806.

(332) Chen, X.; Zhang, P.; Zhang, Y.; Wei, M.; Tian, T.; Zhu, D.; Guan, Y.; Wei, W.; Ma, Y. The research progression of direct NLRP3 inhibitors to treat inflammatory disorders. *Cell. Immunol.* **2024**, *397-398*, 104810, DOI: 10.1016/j.cellimm.2024.104810.

(333) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594, DOI: 10.1021/jm300687e.

(334) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J. Cheminform.* **2015**, *7*, 26, DOI: 10.1186/s13321-015-0078-2.

(335) Patel, H. M.; Sing, B.; Bhardwaj, V.; Palkar, M.; Shaikh, M. S.; Rane, R.; Alwan, W. S.; Gadad, A. K.; Noolvi, M. N.; Karpoormath, R. Design, synthesis and evaluation of small molecule imidazo[2,1-b][1,3,4]thiadiazoles as inhibitors of transforming growth factor-β type-I receptor kinase (ALK5). *Eur. J. Med. Chem.* **2015**, *93*, 599–613, DOI: 10.1016/j.ejmech.2014.09.002.

(336) Ramprasad, J.; Nayak, N.; Dalimba, U.; Yogeeswari, P.; Sriram, D.; Peethambar, S. K.; Achur, R.; Kumar, H. S. S. Synthesis and biological evaluation of new imidazo[2,1-b][1,3,4]thiadiazole-benzimidazole derivatives. *Eur. J. Med. Chem.* **2015**, *95*, 49–63, DOI: 10.1016/j.ejmech.2015.03.024.

(337) Ramprasad, J.; Nayak, N.; Dalimba, U.; Yogeeswari, P.; Sriram, D. Ionic liquid-promoted one-pot synthesis of thiazole–imidazo[2,1-b][1,3,4]thiadiazole hybrids and their antitubercular activity. *MedChemComm* **2016**, *7*, 338–344, DOI: 10.1039/C5MD00346F.

(338) AG, F. H. L. R. Bicyclic ketone compounds and methods of use thereof pat., EP3652178A1, 2020.

(339) AG, F. H. L. R. Bicylic lactams as receptor-interacting protein-1 (RIP1) kinase inhibitors for treating e.g. inflammatory diseases pat., EP3760625A1, 2021.

(340) Seneci, P; Nicola, M; Inglesi, M; Vanotti, E; Resnati, G Synthesis of mono- and disubstituted 1H-imidazo [1,2-B] pyrazoles. *Synth. Commun.* **1999**, *29*, 311–341, DOI: 10.1080/00397919908085772.

(341) Sagar, R.; Mishra, V. K.; Tiwari, G.; Khanna, A.; Tyagi, R. Efficient synthesis of chirally enriched 1H-Imidazo[1,2-b]pyrazole- and 4H-Imidazo[1,2-b][1,2,4]triazole-Based bioactive glycohybrids. *Synthesis* **2024**, *56*, 1017–1025, DOI: 10.1055/a-2157-9100.

(342) Wang, J.-L.; Wu, G.-Y.; Luo, J.-N.; Liu, J.-L.; Zhuo, C.-X. Catalytic intermolecular deoxygenative coupling of carbonyl compounds with alkynes by a Cp*Mo(II)-catalyst. *J. Am. Chem. Soc.* **2024**, *146*, 5605–5613, DOI: 10.1021/jacs.3c14195.

(343) Alanine, T. A.; Galloway, W. R. J. D.; McGuire, T. M.; Spring, D. R. Concise synthesis of substituted quinolizin-4-ones by ring-closing metathesis. *Eur. J. Org. Chem.* **2014**, *2014*, 5767–5776, DOI: 10.1002/ejoc.201402648.

(344) Yu, M.; Lou, S.; Gonzalez-Bobes, F. Ring-closing metathesis in pharmaceutical development: Fundamentals, applications, and future directions. *Org. Process Res. Dev.* **2018**, *22*, 918–946, DOI: 10.1021/acs.oprd.8b00093.

(345) Wieczorek, E.; Sin, J. W.; Holland, M. T. O.; Wilbraham, L.; Perez, V. S.; Bradley, A.; Miketa, D.; Brennan, P. E.; Duarte, F. Transfer learning for Heterocycle Synthesis Prediction. *ChemRxiv* **2024**, This content is a preprint and has not been peer-reviewed., DOI: 10.26434/chemrxiv-2024-ngqqg.

(346) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. "Ring Breaker": Neural network driven synthesis prediction of the ring system chemical space. *J. Med. Chem.* **2020**, *63*, 8791–8808, DOI: 10.1021/acs.jmedchem.9b01919.

(347) Roughley, S. D.; Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011**, *54*, 3451–3479, DOI: 10.1021/jm200187y.

(348) Brown, D. G.; Boström, J. Analysis of past and present synthetic methodologies on medicinal chemistry: Where have all the new reactions gone? *J. Med. Chem.* **2016**, *59*, 4443–4458, DOI: 10.1021/acs.jmedchem.5b01409.

(349) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. In *Advances in Neural Information Processing Systems*, 2017; Vol. 30, DOI: 10.48550/arXiv.1706.03762.

(350) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583, DOI: 10.1021/acscentsci.9b00576.

(351) Hu, W.; Zeng, Q.; Chen, W. Improved synthesis of first cell-permeable allosteric PTPRZ inhibitor NAZ2329. *Russ. J. Gen. Chem.* **2021**, *91*, 2095–2100, DOI: 10.1134/S1070363221100273.

(352) Sahner, J. H.; Empting, M.; Kamal, A.; Weidel, E.; Groh, M.; Börger, C.; Hartmann, R. W. Exploring the chemical space of ureidothiophene-2-carboxylic acids as inhibitors of the quorum sensing enzyme PqsD from Pseudomonas aeruginosa. *Eur. J. Med. Chem.* **2015**, *96*, 14–21, DOI: 10.1016/j.ejmech.2015.04.007.

(353) Zhang, P.; Terefenko, E. A.; Bray, J.; Deecher, D.; Fensome, A.; Harrison, J.; Kim, C.; Koury, E.; Mark, L.; McComas, C. C.; Mugford, C. A.; Trybulski, E. J.; Vu, A. T.; Whiteside, G. T.; Mahaney, P. E. 1- or 3-(3-Amino-2-hydroxy-1-phenyl propyl)-1,3-dihydro-2H-benzimidazol-2-ones: Potent, Selective, and Orally Efficacious Norepinephrine Reuptake Inhibitors. *J. Med. Chem.* **2009**, *52*, 5703–5711, DOI: 10.1021/jm900888c.

(354) Kepp, K. P. A quantitative scale of oxophilicity and thiophilicity. *Inorg. Chem.* **2016**, *55*, 9461–9470, DOI: 10.1021/acs.inorgchem.6b01702.

(355) Sun, W.-W.; Xie, Y.-B.; Deng, T.-T.; Huang, J.; Liu, J.-K.; Wu, B. Acid-promoted cyclization reaction of the guanine base with 1,1,3,3-tetramethoxypropane: A method for the preparation of M1 dG and its derivatives. *Curr. Protoc.* **2023**, *3*, e741, DOI: 10.1002/cpz1.741.

(356) Mercogliano, M.; Iesce, M. R.; Alfieri, M. L.; Buommino, E.; DellaGreca, M. Hands-on synthesis of furanamides and evaluation of their antimicrobial activity. *Nat. Prod. Res.* **2023**, *37*, 3484–3491, DOI: 10.1080/14786419.2022.2087220.

(357) Karabatsos, G. J.; Osborne, C. E. Structural studies by nuclear magnetic resonance – XVI: conformations and configurations of hydrazones. *Tetrahedron* **1968**, *24*, 3361–3368, DOI: 10.1016/S0040-4020(01)92634-1.

(358) Benassi, R.; Benedetti, A.; Taddei, F.; Cappelletti, R.; Nardi, D.; Tajana, A. Conformational analysis of hydrazones. ^{1}H dynamic nuclear magnetic resonance and solvent effects in aryl- and 2-furylaldehyde ethylaminoacetylhydrazones. *Org. Magn. Reson.* **1982**, *20*, 26–30, DOI: 10.1002/mrc.1270200107.

(359) Tatum, L. A.; Su, X.; Aprahamian, I. Simple hydrazone building blocks for complicated functional materials. *Acc. Chem. Res.* **2014**, *47*, 2141–2149.

(360) Su, X.; Aprahamian, I. Switching around two axles: controlling the configuration and conformation of a hydrazone-based switch. *Org. Lett.* **2011**, *13*, 30–33, DOI: 10.1021/ol102422h.

(361) Qian, H.; Pramanik, S.; Aprahamian, I. Photochromic hydrazone switches with extremely long thermal half-lives. *J. Am. Chem. Soc.* **2017**, *139*, 9140–9143, DOI: 10.1021/jacs.7b04993.

(362) Zheng, L.-Q.; Yang, S.; Lan, J.; Gyr, L.; Goubert, G.; Qian, H.; Aprahamian, I.; Zenobi, R. Solution phase and surface photoisomerization of a hydrazone switch

with a long thermal half-life. *J. Am. Chem. Soc.* **2019**, *141*, 17637–17645, DOI: 10.1021/ar500111f.

(363) Sahyoun, T.; Arrault, A.; Schneider, R. Amidoximes and oximes: Synthesis, structure, and their key role as NO donors. *Molecules* **2019**, *24*, 2470, DOI: 10.3390/molecules24132470.

(364) Chandran, N.; Bose, K.; Thekkantavida, A. C.; Thomas, R. R.; Anirudhan, K.; Bindra, S.; Sura, S.; Hasan, H. A.; Kumar, S.; Rangarajan, T. M.; Al-Sehemi, A. G.; Gahtori, P.; Kim, H.; Mathew, B. Oxime Derivatives: A Valid Pharmacophore in Medicinal Chemistry. *ChemistrySelect* **2024**, *9*, e202401726, DOI: https://doi.org/10.1002/slct.202401726.

(365) Resnick, E. et al. Rapid covalent-probe discovery by electrophile-fragment screening. *J. Am. Chem. Soc.* **2019**, *141*, 8951–8968, DOI: 10.1021/jacs.9b02822.

(366) Fagan, V. et al. A chemical probe for Tudor domain protein Spindlin1 to investigate chromatin function. *J. Med. Chem.* **2019**, *62*, 9008–9025, DOI: 10.1021/acs.jmedchem.9b00562.

(367) Wu, Q. et al. A chemical toolbox for the study of bromodomains and epigenetic signaling. *Nat. Commun.* **2019**, *10*, 1915, DOI: 10.1038/s41467-019-09672-2.

(368) Xiong, Y. et al. Discovery of a potent and selective fragment-like inhibitor of methyllysine reader protein spindlin 1 (SPIN1). *J. Med. Chem.* **2019**, *62*, 8996–9007, DOI: 10.1021/acs.jmedchem.9b00522.

(369) Quinlan, R. B. A.; Brennan, P. E. Chemogenomics for drug discovery: clinical molecules from open access chemical probes. *RSC Chem. Biol.* **2021**, *2*, 759–795, DOI: 10.1039/D1CB00016K.

(370) Müller, S. et al. Target 2035 - update on the quest for a probe for every protein. *RSC Med. Chem.* **2022**, *13*, 13–21, DOI: 10.1039/D1MD00228G.

(371) Xiong, Y. et al. Discovery of a potent, selective, and cell-active SPIN1 inhibitor. *J. Med. Chem.* **2024**, *67*, 5837–5853, DOI: 10.1021/acs.jmedchem.4c00121.

(372) Sanfelice, D.; Antolin, A. A.; Crisp, A.; Chen, Y.; Bellenie, B.; Brennan, P. E.; Edwards, A.; Müller, S.; Al-Lazikani, B.; Workman, P. The Chemical Probes Portal - 2024: update on this public resource to support best-practice selection and use of small molecules in biomedical research. *Nucleic Acids Res.* **2025**, *53*, D1663–D1669, DOI: 10.1093/nar/gkae1062.

(373) Frye, S. V. The art of the chemical probe. *Nat. Chem. Biol.* **2010**, *6*, 159–161, DOI: 10.1038/nchembio.296.

(374) Workman, P.; Collins, I. Probing the probes: fitness factors for small molecule tools. *Chem. Biol.* **2010**, *17*, 561–577, DOI: 10.1016/j.chembiol.2010.05.013.

(375) Balıkçı, E. et al. Unexpected noncovalent off-target activity of clinical BTK inhibitors leads to discovery of a dual NUDT5/14 antagonist. *J. Med. Chem.* **2024**, *67*, 7245–7259, DOI: 10.1021/acs.jmedchem.4c00072.

(376) Mildvan, A. S.; Xia, Z; Azurmendi, H. F.; Saraswat, V; Legler, P. M.; Massiah, M. A.; Gabelli, S. B.; Bianchet, M. A.; Kang, L.-W.; Amzel, L. M. Structures and mechanisms of Nudix hydrolases. *Arch. Biochem. Biophys.* **2005**, *433*, 129–143, DOI: 10.1016/j.abb.2004.08.017.

(377) Carreras-Puigvert, J. et al. A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family. *Nat. Commun.* **2017**, *8*, 1541, DOI: 10.1038/s41467-017-01642-w.

(378) Wright, R. H. G. et al. ADP-ribose–derived nuclear ATP synthesis by NUDIX5 is required for chromatin remodeling. *Science* **2016**, *352*, 1221–1225, DOI: 10.1126/science.aad9335.

(379) Page, B. D. G. et al. Targeted NUDT5 inhibitors block hormone signaling in breast cancer cells. *Nat. Commun.* **2018**, *9*, DOI: 10.1038/s41467-017-02293-7.

(380) Pickup, K. E.; Pardow, F.; Carbonell-Caballero, J.; Lioutas, A.; Villanueva-Cañas, J. L.; Wright, R. H. G.; Beato, M. Expression of oncogenic drivers in 3D cell culture depends on nuclear ATP synthesis by NUDT5. *Cancers (Basel)* **2019**, *11*, 1337, DOI: 10.3390/cancers11091337.

(381) Huber, K. V. M. et al. Stereospecific targeting of MTH1 by (S)-crizotinib as an anticancer strategy. *Nature* **2014**, *508*, 222–227, DOI: 10.1038/nature13194.

(382) Zhang, C.-H. et al. From lead to drug candidate: Optimization of 3-(phenylethynyl)-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-amine derivatives as agents for the treatment of triple negative breast cancer. *J. Med. Chem.* **2016**, *59*, 9788–9805, DOI: 10.1021/acs.jmedchem.6b00943.

(383) Sato, K.; Sugimoto, H.; Rikimaru, K.; Imoto, H.; Kamaura, M.; Negoro, N.; Tsujihata, Y.; Miyashita, H.; Odani, T.; Murata, T. Discovery of a novel series of indoline carbamate and indolinylpyrimidine derivatives as potent GPR119 agonists. *Bioorg. Med. Chem.* **2014**, *22*, 1649–1666, DOI: 10.1016/j.bmc.2014.01.028.

(384) Eastwood, P.; Gonzalez, J.; Paredes, S.; Fonquerna, S.; Cardús, A.; Alonso, J. A.; Nueda, A.; Domenech, T.; Reinoso, R. F.; Vidal, B. Discovery of potent and selective bicyclic A(2B) adenosine receptor antagonists via bioisosteric amide replacement. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 1634–1637, DOI: 10.1016/j.bmcl.2010.01.077.

(385) Campuzano, V et al. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **1996**, *271*, 1423–1427, DOI: 10.1126/science.271.5254.1423.

(386) Dürr, A; Cossee, M; Agid, Y; Campuzano, V; Mignard, C; Penet, C; Mandel, J. L.; Brice, A; Koenig, M Clinical and genetic abnormalities in patients with Friedreich's ataxia. *N. Engl. J. Med.* **1996**, *335*, 1169–1175, DOI: 10 . 1056 / NEJM199610173351601.

(387) Delatycki, M. B.; Williamson, R; Forrest, S. M. Friedreich ataxia: an overview. *J. Med. Genet.* **2000**, *37*, 1–8, DOI: 10.1136/jmg.37.1.1.

(388) Pastore, A.; Puccio, H. Frataxin: a protein in search for a function. *J. Neurochem.* **2013**, *126 Suppl 1*, 43–52, DOI: 10.1111/jnc.12220.

(389) Vilema-Enríquez, G.; Quinlan, R.; Kilfeather, P.; Mazzone, R.; Saqlain, S.; Del Molino Del Barrio, I.; Donato, A.; Corda, G.; Li, F.; Vedadi, M.; Németh, A. H.; Brennan, P. E.; Wade-Martins, R. Inhibition of the SUV4-20 H1 histone methyltransferase increases frataxin expression in Friedreich's ataxia patient cells. *J. Biol. Chem.* **2020**, *295*, 17973–17985, DOI: 10.1074/jbc.RA120.015533.

(390) Nishioka, K.; Rice, J. C.; Sarma, K.; Erdjument-Bromage, H.; Werner, J.; Wang, Y.; Chuikov, S.; Valenzuela, P.; Tempst, P.; Steward, R.; Lis, J. T.; Allis, C. D.; Reinberg, D. PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin. *Mol. Cell* **2002**, *9*, 1201–1213, DOI: 10.1016/s1097-2765(02)00548-8.

(391) Yang, H.; Pesavento, J. J.; Starnes, T. W.; Cryderman, D. E.; Wallrath, L. L.; Kelleher, N. L.; Mizzen, C. A. Preferential dimethylation of histone H4 lysine 20 by Suv4-20. *J. Biol. Chem.* **2008**, *283*, 12085–12092, DOI: 10.1074/jbc.M707974200.

(392) Beck, D. B.; Oda, H.; Shen, S. S.; Reinberg, D. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes Dev.* **2012**, *26*, 325–337, DOI: 10.1101/gad.177444.111.

(393) Bromberg, K. D. et al. The SUV4-20 inhibitor A-196 verifies a role for epigenetics in genomic integrity. *Nat. Chem. Biol.* **2017**, *13*, 317–324, DOI: 10.1038/nchembio.2282.

(394) Quinlan, R. Strategies for increasing frataxin expression in Friedreich's ataxia: modulation of the epigenome and proteome, Ph.D. Thesis, University of Oxford, 2021.

(395) Hsiao, K.; Zegzouti, H.; Goueli, S. A. Methyltransferase-Glo: a universal, bioluminescent and homogenous assay for monitoring all classes of methyltransferases. *Epigenomics* **2016**, *8*, 321–339, DOI: 10.2217/epi.15.113.

(396) Gao, K.; Oerlemans, R.; Groves, M. R. Theory and applications of differential scanning fluorimetry in early-stage drug discovery. *Biophys. Rev.* **2020**, *12*, 85–104, DOI: 10.1007/s12551-020-00619-2.

(397) Chen, X. Combining Molecular Modeling and Machine Learning for the Design and Optimization of Small Molecules for Friedreich's Ataxia, MA thesis, University of Oxford, 2024.

(398) Lanman, B. A. et al. Discovery of a covalent inhibitor of KRASG12C (AMG 510) for the treatment of solid tumors. *J. Med. Chem.* **2020**, *63*, 52–65, DOI: 10.1021/acs.jmedchem.9b01180.

(399) (US), G. W. I. Amino acid derivatives as NO Synthase inhibitors. Pat., US5874472A, 1999.

(400) Landry, M. L.; Crawford, J. J. LogD contributions of substituents commonly used in medicinal chemistry. *ACS Med. Chem. Lett.* **2020**, *11*, 72–76, DOI: 10.1021/acsmedchemlett.9b00489.

(401) Sheriff, S. et al. Small molecule receptor protein tyrosine phosphatase γ (RPTPγ) ligands that inhibit phosphatase activity via perturbation of the tryptophan-proline-

aspartate (WPD) loop. *J. Med. Chem.* **2011**, *54*, 6548–6562, DOI: 10 . 1021 / jm2003766.

(402) Davies, H. M.; Calvo, R. L.; Townsend, R. J.; Ren, P; Churchill, R. M. An exploratory study of type II [3 + 4] cycloadditions between vinylcarbenoids and dienes. *J. Org. Chem.* **2000**, *65*, 4261–4268, DOI: 10.1021/jo991959b.

(403) Hawker, D. D.; Silverman, R. B. Synthesis and evaluation of novel heteroaromatic substrates of GABA aminotransferase. *Bioorg. Med. Chem.* **2012**, *20*, 5763–5773, DOI: 10.1016/j.bmc.2012.08.009.

(404) Muguruma, H.; Saito, T.; Sasaki, S.; Hotta, S.; Karube, I. Synthesis and characterization of α,α'-bis(aminomethyl)oligothiophenes and their related compounds. *J. Heterocycl. Chem.* **1996**, *33*, 173–178, DOI: 10.1002/jhet.5570330130.

O Lord God, when thou givest to thy servants to endeavour any great matter, grant us also to know that it is not the beginning, but the continuing of the same, until it be thoroughly finished, which yieldeth the true glory; through him who for the finishing of thy work laid down his life for us, our Redeemer, Jesus Christ. Amen.

Sir Francis Drake (1540-1596)