DISS. ETH NO. 26293

FILLING IN THE GAPS: GEOMETRIC COMPLEMENTARITY AND ITS ROLE IN LIGAND-PROTEIN INTERACTION PREDICTION

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

RYAN BYRNE

MSc.,
Imperial College London

MPharm.,
Queen's University, Belfast

born on 19.04.1991

citizen of
United Kingdom of Great Britain and Northern Ireland
Republic of Ireland

accepted on the recommendation of

Prof. Dr. Gisbert Schneider
Prof. Dr. Jonathan Hall
Prof. Dr. Donald Hilvert

2020

'Either this is madness or it is Hell'.'It is neither', calmly replied the voice of the Sphere, 'it is Knowledge; it is Three Dimensions: open your eye once again and try to look steadily.'

Edwin Abbott, *Flatland: A Romance of Many Dimensions*

Abstract

The identification of small-molecule inhibitors of macro-molecular targets is a challenging task, and one to which considerable resources have been devoted over the preceding decades. Huge compound libraries have been synthesised and trialled, and yet still capture only a tiny fraction of the available chemical space. At a fundamental level, we are interested in learning how chemical and biological spaces interrelate, to describe what makes a ligand, or a target, promiscuous, and what makes some targets 'harder' than others to ligand. As such, methods which allow us to gain a broader perspective are valuable, if they can guide us in new directions.

One compelling conceptual framework which has been proposed to describe the process underlying these relations is that of shape complementarity. This has the dual advantages of an intuitive relation to the induced-fit and conformational-selection models of ligand-target interaction, and substantial support from experimental data. Shape-based approaches have been used to identify novel compounds, representing 'scaffold hops' into new chemical space. These hops are immensely valuable, on both a scientific and a commercial basis, representing genuinely novel chemical insights into biological problems. Shape-based screening is not without issues; the results can be counter-intuitive, the definition is subject to debate, the good methods are often slow, and the fast methods give at best a rough sketch of a molecule, rather than a finely detailed likeness.

Shape-based approaches which help us to 'look before we scaffold hop', to encompass a wider view of chemical space, which capture shape in an intuitive fashion, and do so in a reasonable time-frame, are of considerable value in the pursuit of new chemical perspectives on biology.

Equally, the shape-complementarity hypothesis motivates the development of methods which allow us to compare the geometries of binding sites, and then to assess the relationship between ligand and target shape.

Here, we develop and describe a novel method for the generation of global shape descriptors and shape-based fingerprints, capturing these through the local fractal dimension of surface representations of small- and macro-molecules. Given the profusion of descriptors available to the computational chemist, we focused our initial efforts on demonstrating the novelty, validity, and utility of the developed method.

We adopted a ligand-based retrospective benchmarking platform, and used this to compare our approach with existing shape-based and simple methods. We saw that our method recovered more, and more diverse, actives for a set based on medicinal chemical efforts than a gold-standard commercial technology, in a thousandth of the time. Its performance on other sets was broadly equivalent. We used data extracted from this experiment to train probabilistic models to determine the form of the relationship between chemical and biological similarity for the methods profiled.

Given this promising early result, we conducted two prospective screening rounds with the developed methods. The first, a small screen and subsequent SAR study based on the natural product (-)-Englerin A, utilised the global shape descriptor to identify a novel inhibitor with no known analogues for the described target. The second, utilising the local shape fingerprint, was conducted against seven targets. 22% (n=28) of all screened compounds were active against the desired targets. These contain 22 scaffolds which were not previously publicly-known to modulate the targets. As such, the retrieved compounds represent novel chemical space for the targets of interest. Additionally, we assessed the performance of the trained probabilistic model, and found the hit rate at the chosen cutoff to be 42%, which captures 93% of the hit compounds. The probability model based on simple similarity was found to have higher predictive power than more advanced methods for this set of compounds.

To explore the relationship between ligand and target shape, one of the hit compounds identified was profiled in depth, determining its IC_{50} against the Pim-1 kinase (484 nM), subjecting it to a kinase selectivity panel, and obtaining a crystal structure. It was discovered to be a highly-selective inhibitor, potently inhibiting the Pim-1, GSK-3β and DRAK-1 kinases. Analysis of 133 existing crystal structures revealed that the compound binds in a similar fashion to known inhibitors, with few specific contacts. The advantages of a shape-based method which is less sensitive to conformation are highlighted by comparison with ROCS. There is some correlation between protein and ligand shape-fingerprint similarity.

Having shown that a shape-based method suffices to identify active compounds for a set of targets, we sought to determine the feasibility of reversing the relationship, suggesting ligand shape fingerprints based on the shape of a binding pocket. To do this, we analysed fourteen thousand published crystal structures, calculating ligand and pocket shape-fingerprints, and using a portion of these to train a sequence transduction model. Approximately a fifth of the reconstructed test ligand fingerprints were highly-similar to those extracted from the crystal structure, and all were significantly better than a simple statistical model.

In summary, this work seeks to demonstrate the novelty and validity of the described approach for capturing shape information, and the utility of such approaches in the description of chemical space, both small- and macro-molecular.

Zusammenfassung

Die Identifizierung von niedermolekularen Inhibitoren für makromolekulare Zielproteine ist eine herausfordernde Aufgabe und eine, für die über viele Jahrzehnte erhebliche Mittel zur Verfügung gestellt wurden. Grosse Substanzbibliotheken wurden synthetisiert und gegen Zielproteine getestet, jedoch stellen diese nur winzige Anteile des erreichbaren Chemical Space dar. Davon ausgehend sind Methoden wertvoll, die uns erlauben eine Perspektive für einen breiteren Bereich des zugrunde liegenden Chemical Space zu kriegen, vor allem, wenn sie uns in eine neue Richtung führen. Auf einer grundsätzlichen Ebene sind wir daran interessiert zu verstehen wie sich der chemische und der biologische Raum überschneiden, zu beschreiben was einen Liganden oder ein Zielprotein mehrere Partner haben lässt, und zu erkennen was es für bestimmte Zielproteine herausfordernder macht Liganden zu finden als für andere.

Ein Rahmenkonzept, das den Prozess von diesen zugrunde liegenden Verhältnissen beschreibt, ist das der Shape-Komplementarität. Dessen dualer Vorteil besteht in eine intuitiven Beziehung für Liganden-Zielprotein Interaktionen zwischen dem Induced-fit und Modellen zur Konformationsauswahl, sowie einer starken Untermauerung basierend auf einer Fülle an experimenteller Daten. Shape-basierten Screenings sind nicht ohne Herausforderungen; Das Resultat kann entgegen der Intuition sein, deren Definition kann Gegenstand der Debatte sein, gute Methoden sind häufig langsam, und die Bestleistung schneller Methoden ist, eher ein grober Umriss des Moleküls, als ein voll detailliertes Abbild dessen.

Shape-basierende Ansätze, die uns helfen, bevor wir einen Scaffold-hop machen, einen größeren Überblick im Chemical Space zu kriegen, sollen die Kontur eines Moleküls auf eine intuitive Art einfangen, und dies in einem nützlichen Zeitrahmen tun. Erfüllt ein Ansatz dies, kreiert es einen beachtlichen Wert, wenn es darum geht neue chemische Perspektiven auf die Biologie zu bekommen. Genauso motiviert uns die Hypothese der Shape-Komplementarität bei der Entwicklung von Methoden, die uns erlauben Bindungsseitengeometrien zu vergleichen, und dadurch die Beziehung zwischen Liganden-Shape und Zielprotein-Shape zu messen.

In dieser Arbeit entwickeln und beschreiben wir eine neue Methode, um globale Shape Deskriptoren und shape-basierte Fingerprints zu generieren, die über die lokale fraktale Dimension der Oberflächendarstellung von niedermolekularen und makromolekularen Verbindungen erfasst wird. Aufgrund der Fülle an Deskriptoren, die einem Computerchemiker zur Verfügung stehen, fokussieren wir unser in erster Linie darauf Neuartigkeit, Validität und Nützlichkeit der entwickelten Methode aufzuzeigen.

Wir nehmen eine liganden-basierte, retrospektive Benchmarking Plattform und nutzen diese um unseren Ansatz mit existierenden shape-basierten und einfacheren Methoden zu vergleichen. Basierend auf einem Set von Verbindungen mit medizinal-chemisch Ursprung, hat unsere Methode mehr und diversere, aktive Verbindungen gefunden, als eine kommerzielle Technologie, die als Goldstandard im Feld gilt, und das Ganze in einem tausendstel der Zeit. Die Leistung unserer Methode auf andere Sets war weitgehend gleich. Wir haben, von diesem Experiment stammende, Daten genutzt um Wahrscheinlichkeits-Modelle zu trainieren, die bestimmen, welche Form der Beziehung zwischen chemischer und biologischer Ähnlichkeit für die verwendeten Methoden besteht.

Aufgrund dieser vielversprechenden frühen Resultate haben wir zwei prospektive Screeningrunden mit den entwickelten Methoden durchgeführt. Das erste, kleinere Screening und die darauffolgenden SAR Studie, basierend auf dem Naturstoff (-)-Englerin A, nutzte den globalen Shape-Deskriptor, um neue Inhibitoren zu identifizieren, von denen noch keine Analoga für die beschriebenen Zielproteine getestet wurden. Das zweite Screening umfasste sieben Zielproteine und nutzte den lokalen Shape-Fingerprint. In diesem Screening sind 22% (*n*=28) der Verbindungen aktiv gegen das jeweilige Zielprotein und umfassen 22 Molekülgerüste, deren Aktivität gegen die Zielproteine bisher nicht bekannt waren. Als solche stellen die gefundenen Verbindungen neuen Chemical Space für die Zielproteine dar. Zusätzlich wurde die Leistung des trainierten probabilistischen Modells bewertet und die Hitrate, bei dem gewählten Schwellenwert, ist 42%, womit 93% der Hits erfasst werden. Das Wahrscheinlichkeitsmodell basiert auf einfacher Ähnlichkeit und hat, für dieses Set an Verbindungen, eine höhere Vorhersagekraft als komplexere Methoden.

Um die Beziehung zwischen der Kontur von Liganden und Zielproteinen zu erforschen, wurde einer der identifizierten Hitverbindungen weiter untersucht. Dessen IC_{50} gegen die Pim-1 Kinase (484 nM), und seine Selektivität gegen andere Kinasen, mittels einer Kinase-Selektivitäts-Auswahl, wurden bestimmt und seine Kristallstruktur erhatlen. Diese Verbindung ist ein hoch-selektiver Inhibitor mit potentieller Inhibition von Pim-1, GSK-3β und DRAK-1 Kinasen. Die Analyse von 133 existierenden Kristallstrukturen zeigte, dass die Verbindung eine ähnliche Bindungsstrategie, mit wenig spezifischen Kontaktpunkten, wie andere bekannte Inhibitoren hat. Durch den Vergleich mit ROCS wird der Vorteil zum Vorschein gebracht, dass die Shape-basierten Methode weniger beeinflussbar durch Strukturkonformation ist.

Wir haben gezeigt, dass eine Shape-basierte Methode ausreicht um aktive Verbindungen für ein Set von Zielproteinen zu identifizieren. Wir untersuchten die Machbarkeit, die Beziehung zwischen Ligand und Bindungstasche umzudrehen, und damit zu suggerieren, dass Shape-Fingerprints von Liganden basierend auf der Bindungstaschen-Kontur bestimmt werden können. Um dies zu tun, analysierten wir vierzehntausend publizierte Kristallstrukturen, berechneten deren Liganden- und Bindungstaschen-Shape-Fingerprint, und nutzten das Verhältnis dieser, um ein Sequenz-Transduktion-Modell zu trainieren. Ungefähr ein Fünftel der rekonstruierten Testliganden Fingerprints war hochgradig ähnlich zu jenen, die von den Kristallstrukturen extrahiert wurden, und alle waren signifikant besser, als ein simples statistisches Modell.

Zusammenfassend hat diese Arbeit die Neuartigkeit und Validität des beschriebenen Ansatz, um Shape Information zu erfassen, und die Nützlichkeit eines solchen Ansatz, um den Chemical Space für niedermolekulare und makromolekulare Verbindungen zu beschreiben, demonstriert.

Acknowledgements

The past few years have been somewhat eventful, to put it mildly.

I would like to thank Gisbert for giving me the opportunity to spend some time in the lab. To a profound degree, your constant creativity, scientific discussions and the occasional non sequitur have kept me going, and mostly on the right path. As you put it, I think we've learned a lot from one another, and most of it good.

In no particular order, I'd like to thank my labmates and friends from the past few years.

Alex Button, for the consistently stimulating and energetic conversation, from my first days in the lab until the present, and for your creative approach to problem-solving. Berend Huisman, for introducing me to the splendour and complexity of Dutch food, and for your consistently kicking my feet under the table, as a reminder to stay in the present. Erik Gawehn, for your approach to life. Gisela Gabernet Garriga, for your good humour and guidance in my early days. Dominique Bruns, for your eternally-sunny disposition, extreme kindness, and love of papier-mâché mountains. Claudia Neuhaus, for your friendship, consideration, and love of irritating board games. Alex Müller, for demonstrating (at length) that the Swiss can be cool, and that I'll never be fit so I should give up now.

Lukas Friedrich, for the companionship, collaboration, and your unique taste in music. Xuejin (JJ) Zhang, for your kindness, for rejecting the boring, and for your refreshing outlook on life. Christoph Bauer, for my new understanding of the state of affairs in Austrian academia, and for your sunny disposition. Francesca Grisoni, for the many entertaining and informative discussions, and the occasional lecture on descriptors, MATLAB, and descriptors in MATLAB. Jens Fuchs, for the time spent together, for being an all-round good guy, and for your approach to life. Cyril Brunner, for your friendship, innocence, and for organising my life better than I ever could. Damian Gautschi, for bringing me to a terrible club that I hated, and for all the conversations over the years. Michael Moret, for introducing me to the finer points of French society, for arguing with ticket conductors, and for solidarity in shared scientific suffering. Daniel Merk, for your assistance, collaboration, and remarkable drive.

I'd like to thank my students, and those who have been in and out of the lab over the years, who have been wonderful: Veronika Bobinger, Eman Darwish, Reto Höhener, Mattis Hilleke, Alice Lessing, Benedikt Winkler, Tatu Lindroos, Elena Gelžintye, Anvita Gupta, Emanuele Rossi, Moritz Gück, Stefan Gugler, Arpad Dunai, Joana Sigrist, Robin Lingwood, and Shinji Iida. Any omissions are a mark of my advanced age. Niamh O'Neill, Tao Jinyan, and Aakriti Sethi, for tolerating my entirely inadequate supervision with humour and patience, and for teaching me how to teach better. Aakriti, your words will stick with me, always.

My heartfelt gratitude to the Marie Skłodowska-Curie actions for providing the ESR programme, which facilitated my time at the ETH, for the interesting courses and opportunities, and, more than anything, for introducing me to a great bunch of people across Europe. To Charlotte, Roberto, Francesca, Engi, Atilio, Giulia, Patrick, Laura, Valeria, Simon (João), Joy, Markella, Ave, and Maxime: for forming a nucleus of support, and for the good times across Europe, I'm forever grateful.

For the BigChemists in the frozen North (Mike, Thomas, Ollie, Laurianne, and Josep) - thank you for making my sojourn in Sweden so pleasant. A special thanks to Drs. Hongming Chen and Ola Engkvist for hosting me there.

I would like to thank Switzerland (as a nation) for providing a home for a rather confused British person, at a rather confusing time, and showing me how well things can work when people trust one another. On that note, cheers to Camille Glaus and Gabriel Fiette, for the constitutional discussions and introduction to life beyond the Röstigraben.

Finally, I would like to thank my family and friends, for their ongoing support. Even though we are dispersed, we still come together when we need to.

Acronyms

AAE Adversarial Autoencoder.

ACPC AutoCorrelation of Partial Charges.

Adam adaptive Momentum.

ADMET absorption, distribution, metabolism, excretion and toxicity.

AFD atomistic FD.

AIAYN Attention is all you need.

AJD α-shape Joint-Density.

ANN Artificial Neural Networks.

AUC Area Under Curve.

AVA all-vs.-all.

BEDROC Boltzmann-Enhanced Discrimination of ROC.

BFS breadth-first search.

BMS Bemis-Murcko Scaffold.

BMSE Bemis-Murcko scaffold enrichment.

CATS Chemically-Advanced Template Search.

CNN Convolutional Neural Networks.

CoMFA Comparative Molecular Field Analysis.

CoMMA Comparative Molecular Moment Analysis.

CoMSIA Comparative Molecular Similarity Indices Analysis.

DNN Deep Neural Network.

DOGS Design Of Genuine Structures.

DUD Directory of Useful Decoys.

DUD-E Enhanced DUD.

E3FP Extended Three Dimensional FingerPrint.

ECFP Extended-Connectivity FingerPrint.

EF Enrichment Factor.

FD Fractal Dimensionality.

FDFP Fractal Dimensionality FingerPrint.

GFD global FD.

GMS Generalised BMS.

GMSE generalised BMS enrichment.

GRU gated recurrent unit.

HTS High-Throughput Screening.

IUPAC International Union of Pure and Applied Chemistry.

LBVS Ligand-Based Virtual Screening.

LoRI Local Roughness Indicator.

LSTM long short-term memory.

MACCS Molecular ACCess System.

MCC Matthews correlation coefficient.

MDDR MDL Drug Data Report.

MIF Molecular Interaction Fields.

MIMICS Machine-based Identification of Molecules Inside Characterized Space.

MUV Maximum Unbiased Validation.

MWU Mann-Whitney U.

NCE new chemical entities.

OVA one-vs.-all.

OVO one-vs.-one.

PMI Principal Moments of Inertia.

QSAR Quantitative Structure-Activity Relationship.

REPROVIS-DB REProducible VIrtual Screens DataBase.

RIE Robust Initial Enhancement.

RMSD Root-Mean Square Deviation.

RNN Recurrent Neural Networks.

ROC Receiver Operating Characteristic.

ROCS Rapid Overlay of Chemical Structures.

RS Rank Score.

SABRE Shape-Approach-Based Routines Enhanced.

SAS Solvent-Accessible Surface.

SEA Similarity Ensemble Approach.

SEF Scaffold Enrichment Factor.

SES Solvent-Excluded Surface.

SGD stochastic gradient descent.

SH Spherical Harmonics.

SHAEP Shape and Electrostatic Potential.

SHAFTS SHApe-FeaTure Similarity.

SMILES simplified molecular input line entry specification.

SPiDER Self-organizing map–based Prediction of Drug Equivalence Relationships.

SuMo Surfing the Molecules.

Tc Tanimoto coefficient.

TRPC transient receptor potential canonical.

TRPM transient receptor potential melatstatin.

USR Ultrafast Shape Recognition.

VAE Variational Autoencoder.

WEGA WEighted Gaussian Algorithm.

WHIM Weighted Holistic Invariant Molecular.

WS Window Score.

Contents

Abstract iii

Zusammenfassung vii

Acknowledgements xi

Acronyms xiii

Contents xvii

1 Introduction 1
- 1.1 Computational Chemistry - Similarity and Repetition . . . 1
- 1.2 Describing the undefined - Molecular similarity 4
  - 1.2.1 Molecular properties to substructures - 0D/1D . . . 4
  - 1.2.2 Molecular Fingerprints - 2D 4
  - 1.2.3 Molecular Geometries - 3D 5
  - 1.2.4 Alignment-dependent methods 5
  - 1.2.5 Alignment-independent methods 10
  - 1.2.6 Target similarity 14
  - 1.2.7 Fractals and fractal dimension 18
- 1.3 Molecular similarity comparison 27
- 1.4 Benchmarking 31
- 1.5 Applications 36
  - 1.5.1 Prospective screening 36
  - 1.5.2 Deep learning in shape-based drug discovery . . . 38
  - 1.5.3 Applications in generation and regeneration 48

2 Aims 57

3 Methods **59**
3.1 FRACTVS Package . . . 59
3.1.1 calculate_correlation_exponents.py . . . 63
3.1.2 molecular_representations.py . . . 64
3.1.3 Washing procedure . . . 67
3.1.4 Conformer generation . . . 68
3.1.5 fractal_calculators.py . . . 71
3.1.6 fingerprinter.py and graph_construction.py . . . 74
3.1.7 PDBProcess.py . . . 76
3.2 Benchmarking study . . . 77
3.2.1 Datasets . . . 77
3.2.2 Benchmarking study workflow . . . 79
3.2.3 Ligand-based screening methods . . . 79
3.2.4 Similarity Ranking . . . 84
3.2.5 Performance metrics . . . 86
3.2.6 Benchmarking study analysis . . . 88
3.3 Shape searching for de novo (-)-Englerin A Mimetics . . . 94
3.3.1 Biological assay . . . 96
3.3.2 Analysis of results . . . 96
3.4 Prospective study . . . 97
3.4.1 Screening . . . 97
3.4.2 Analysis . . . 99
3.5 PIM1 - detailed analysis . . . 101
3.5.1 Selectivity study . . . 101
3.5.2 PIM1 - Inhibitor crystal structure . . . 104
3.6 Protein FD - Back Translation and Analysis . . . 106
3.6.1 Dataset construction . . . 106
3.6.2 Models . . . 106
3.6.3 Performance assessment . . . 109
3.7 Computational resources . . . 111

4 Results and Discussion **113**
4.1 Benchmarking study . . . 113
4.1.1 Retrospective Virtual Screening Performance . . . 114
4.1.2 Distribution and probabilistic analysis . . . 129
4.1.3 Algorithm Run-time Performance . . . 135
4.2 Shape searching for de novo (-)-Englerin A Mimetics . . . 138
4.2.1 Similarity and SAR studies . . . 138
4.2.2 Comparative similarity analysis . . . 142
4.2.3 Bioactivity Results . . . 144
4.2.4 Discussion . . . 145

4.3 Prospective study 147
4.3.1 Target overview 148
4.3.2 Compound selection 149
4.3.3 Probability Models 176
4.4 Shape Screening for PIM1 Inhibitors 181
4.4.1 Activity data 181
4.4.2 Selectivity analysis 182
4.4.3 Crystallographic study 189
4.4.4 Discussion 201
4.5 Back-translation study 204
4.5.1 Results 204
4.5.2 Discussion 207

5 Conclusion **215**

Bibliography **221**

Appendix **261**
Englerin study 261
Predicted active compounds 261
Bioactivity determination 261
PIM1 Study 266
Kinase Selectivity 266
Crystallographic parameters 270

Chapter 1

Introduction

1.1 Computational Chemistry - Similarity and Repetition

The primary task of the computational chemist is to impose order on the chemical world, to help us better rationalise our decision-making when it comes to which compounds might work for a given problem, and why. More specifically, they are concerned with how best to arrange, and subsequently search through, the vastness of chemical space in a structured manner, so as to optimise for a few separate objectives simultaneously. Estimates for the size of said space vary widely, from 10^{23} [1] to 10^{180} [2] compounds, and many in-between [3], depending on the definition chosen. Even limiting this to very small molecules, a virtually-enumerated library of all chemically-reasonable combinations of 17 heavy atoms has 166 billion entries [4]. Naturally, much of this space is of no immediately-obvious interest to medicinal chemistry, and compound libraries accumulated to date for High-Throughput Screening (HTS) by various pharmaceutical companies typically have counts in the low millions [5], representing a tiny fraction of accessible chemical matter, whilst achieving substantial successes across a wide range of target families [6]. Plans to increase this by an order-of-magnitude or more [7], still leave much of the chemical world in darkness.

This presents two issues; first, the cost, and sometimes questionable utility of this approach. Typical validated hit-rates for HTS screens are between 0.01 and 0.14% of screened compounds, for screening panels of between one-hundred thousand and a million compounds [8–10]. While screening large numbers of compounds more efficiently [11–13] can help to reduce costs, it does not necessarily improve the validated hit rate. Attempts have been made [14, 15] to rationalise which subsets of a compound library to choose first, to maximise potential benefit, as have efforts to improve the libraries themselves [16]. Whilst early-stage efforts are comparatively inexpensive compared to later stages of the drug development process, a lot of value is created early on, and good quality lead structures can improve the odds of a compound getting through later stages of the process. As the long decline in pharmaceutical productivity continues, promoting consistent efforts to rationalise each stage of the process allows for the early identification of problem candidates, reducing overall attrition rates [17, 18].

Secondly, the titanic effort made to date only represents a tiny fraction of chemical space. Surprisingly, for those cases where pharmaceutical libraries have been compared between companies, the overlap ranged between 1.5% [5] and 10% [19, 20]. As with the combinatorial chemistry approaches much in vogue in earlier years, promotion of diversity is a strong focus [21, 22] in library construction [23] However, defining chemical diversity is, in itself, problematic, and often context- and target-dependent [24]. Commonly, it is used to refer to the variation in scaffolds, the molecular substructure underlying common drug classes [25, 26], or, in medicinal chemistry, to describe an approach developing sets of reactions and substituents which allow for easy reach into different areas of chemical space [27, 28]. The importance of developing new chemotypes, and of successfully emulating natural product compounds with medicinal chemical efforts [29, 30], has led to somewhat philosophical discussions of the correct route for synthetic efforts to take [31, 32].

Fundamentally, chemical space is vast, and not intrinsically ordered at a macro-scale in a way which reflects biological observations. Compounds which are superficially similar can have drastically different on-target effects, diminishing, or inverting, the observed biochemical or phenotypic response [33, 34]. Equally, dissimilar compounds can affect the same target. Complicating matters further are the issues of compound (and target) promiscuity, and polypharmacology. A single (promiscuous) compound often interacts with multiple targets, and vice versa [35]. This is not inherently problematic [36–39], but does complicate the rosy mental image of neatly-divided fields in the chemical landscape.

If we wish to take a pragmatic approach to defining molecular similarity, then, there are a few fundamental questions which we should answer:

1. How can we describe small- and macro-molecules?
2. How can we compare those descriptions?
3. How confident can we be that those comparisons are useful?
4. What can we do with them?

1.2 Describing the undefined - Molecular similarity

To define the relationships between compounds in this vastness, then, in a computationally-feasible, and pharmacologically-meaningful fashion, we require some means of describing our compounds which is both computer-legible and biochemically-relevant in the first instance. Early efforts at developing such means were primarily focused on quantifying small, rational changes to molecular structure, with the aim of describing the relationship between such changes and the effect of the molecule, the Quantitative Structure-Activity Relationship (QSAR) approach [40]. This was facilitated through the development of 'descriptors', numerical encodings of chemical properties, although simple grammatical descriptions, such as the SMILES system for representing small molecules as a text-string, have grown increasingly popular as novel means of utilising them have been developed.

1.2.1 Molecular properties to substructures - 0D/1D

Remaining with the former category for now, descriptors vary substantially in size, theoretical complexity, and intended usage [41]. At their simplest (0D) level, they encode experimental or simple molecular properties, such as weight, or LogP, of an entire molecule. The next level (1D) incorporates information on substructures, such as Molecular ACCess System (MACCS) keys [42] as either binary (absent/present) or frequency vectors.

1.2.2 Molecular Fingerprints - 2D

These approaches are useful, and correlate well with chemical understanding. As the dimensionality considered increases, typically we find a reduction in comprehensibility, and, ideally, a commensurate, compensating effect in efficacy, although this is highly task-dependent. "2D" descriptors, such as the popular Daylight and Extended-Connectivity FingerPrint (ECFP) [43] (the latter of which will be discussed in greater detail in subsection 3.1.6), attempt to capture information about atoms, and their bonds, in a graph-based approach. Typically, these approaches label atoms with atom-specific information (e.g. valency, charge, weight), and the considers local environments of such labels, building these into overall representations of the molecules. They differ from the 1D approaches in their flexibility, as the groups are not pre-defined, and in the process used to convert these environments into the eventual fingerprint representation. These approaches generally allow one to specify a radius (ECFP) or path-length (RDK, Daylight) parameter, which defines the scale at which we wish to consider the molecule's features. Another form of 2D descriptor are pharmacophoric vectors, such as the Chemically-Advanced Template Search (CATS) [44, 45] descriptor, which describes atoms in terms of annotated pharmacophoric features, and builds a vector representation on their cross-correlation in topological distances. This approach attempts to capture the "ensemble of steric and electronic features that are necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" (as defined by International Union of Pure and Applied Chemistry (IUPAC) [46]), in a framework which allows for rapid comparisons over target sets.

1.2.3 Molecular Geometries - 3D

Geometric (3D) approaches attempt to capture more of the physical reality of a ligand than lower levels, by taking into account the coordinates of energy-minimised molecular representations. As this subject is the primary focus of this work, we will expand on this matter in greater depth than for the lower-dimensional representations. These methods can be divided into two rough categories; alignment-dependent, and alignment-independent methods. The former, including Shape and Electrostatic Potential (SHAEP) [47], and the popular Rapid Overlay of Chemical Structures (ROCS) [48] family of approaches, are not necessarily what would commonly be termed as descriptors, as, while they do compare vectorised molecular representations, these must be altered for each pair of compounds considered, rather than being invariant.

1.2.4 Alignment-dependent methods

Gaussian approaches - ROCS, WEGA

With the alignment-based approaches, the underlying principle is that molecules with similar distributions of volume in space should have similar shapes [49]. At its simplest, Root-Mean Square Deviation (RMSD) could be viewed as a simple descriptor of alignment between molecules, although it does not generalise easily to non-identical molecules (this can be accomplished with flexible atom-mapping approaches, however [50]).

Moving from that approach, and incorporating volumetric information, we have hard-sphere and Gaussian-sphere representations, which do not require explicit mapping, and can incorporate relevant information such as atom size, which is correlated with pharmacophore types to a certain extent. The hard-sphere approaches, treating molecules as collections of intersecting spheres of radius equal to their van der Waals radii for example, were initially developed as part of an effort to improve volumetric calculations for macromolecules [51, 52], before their applications in small molecule comparison were considered [53]. In the latter work, which also introduced atom-typing, directly incorporating pharmacophoric characteristics, they introduce the necessary mathematics to facilitate rigid optimisation (rotation and translation) such as to maximise the overlap between any given pair of conformations. In essence, this approach considers similarity as the ratio of the of the intersection of two volumes with their union, and seeks to find the best spatial configuration to maximise this quantity. The pharmacophoric model is implemented by considering sphere 'type', inferred from atomic characteristics, when conducting volume-matching. An alternative form, based on finite-thickness molecular 'skins' was also proposed, which aimed to correct for some of the distortions incumbent upon volume-overlap approaches, and to facilitate ligand-macromolecule analysis, was also proposed [54]. Importantly, these works recognised that alignment-based approaches can become stuck in local overlap maxima, and so introduced a search strategy to minimise the impact of such events. However, this approach was not capable of flexibly matching one conformation onto another, and took approximately 96 wall-clock hours (1993, Silicon Graphics R4000-50) for 192 pairwise searches.

An elegant extension of this approach was proposed, incorporating a 'soft', Gaussian description of the atomic spheres [55, 56] model. This approach, which later became best known, after further development, as ROCS, was based on insights garnered from early work on aligning *ab initio* electron density maps for the purposes of shape comparison [57–59], and the observation that equations of a similar form had been used extensively in quantum chemistry [60] and in the analysis of solvation behaviours [61]. It benefits from offering a logical means of avoiding the computational complexity of calculating the overlap of n-spheres, for which analytical derivatives were unavailable , necessitating expensive refactoring, and from being closer, conceptually, to a 'true' representation of molecular surface than the hard-sphere model. The primary advantages, however, are mathematical and computational, as expressing volumes as Gaussians allowed for easy integration and differentiation (and consequently facilitated the process of alignment enormously).

We will briefly discuss the mathematics facilitating this, adapted from the first ROCS-specific publication [48]. This overlap of Gaussians approach defines shape as;

$$O_{\mathrm{A,B}}\left(\vec{q}^{\mathrm{A}},\vec{q}^{\mathrm{B}}\right) = \iiint \chi^{\mathrm{A}}\left(\vec{r},\vec{q}^{\mathrm{A}}\right)\chi^{\mathrm{B}}\left(\vec{r},\vec{q}^{\mathrm{B}}\right)\mathrm{d}\vec{r} \tag{1.1}$$

where the triple integral indicates integration over $(\mathrm{d}\vec{r} = \mathrm{d}x\mathrm{d}y\mathrm{d}z)$, r is a position in space, q some set of variables defining orientation and position, and χ represents a 'characteristic volume' function. The primary difference is in the form of χ ,where it is defined simply as an inclusion (hard boundary, Heaviside) measure for the hard-sphere case, and as

$$\chi(\vec{r}) = 1 - \prod_{l=1}^{i=N}\left(1 - g_i(\vec{r})\right) \tag{1.2}$$

For the Gaussian extension of the concept, where

$$g_i(\vec{r}) = p_i\mathrm{e}^{-\gamma\vec{r}_i^2} \tag{1.3}$$

And the local co-ordinate centre is defined as $r_i = |\vec{r}_i| = |r - \vec{c}_i|$, from the atomic centre, $(\vec{c}_i)$. An atom, of radius R, is included as a Gaussian of width γ, according to

$$\gamma = \pi\left(\frac{3p}{4\pi R^3}\right)^{2/3} \tag{1.4}$$

Where γ simply allows one to approximate the corresponding volume of the hard-sphere volume. Owing to the Gaussian product rule, where the product of n-many Gaussians can be expressed as a finite sum of the distributions, situated at a point on their original axes, integration of the product of the Gaussian representations is radically-easier than for the hard-body, discontinuous case. In practice, only relations up to the sixth order are considered. The interested reader is referred to the original papers for lengthier discussion of the rationale behind the initial formulation [55] and early applications [56]. A further optimisation to the search heuristic mentioned earlier is given, which involves alignment of two molecules along the axes of their moments of inertia, and choosing starting locations by rotating the starting position by π radians in each axis. This approach will be referred to as ROCS Shape hereafter.

Issues with the implementations of the Gaussian models were explored in recent work [62], highlighting the tendency of such methods to overestimate similarity for 'crowded' atoms, meaning, in essence, that they would over-weight ring structures in their alignment. Their solution, the WEighted Gaussian Algorithm (WEGA), adds a corrective factor, weighting the contribution of each Gaussian term in accordance with how 'crowded' its local neighbourhood is. In addition, this research group published a parallelisation of their procedure [63], which can process 110 million conformations in two seconds using a combination of clever heuristics, and massive computational power; their approach runs on the Tianhe-2 supercomputer, and uses 80 GPU nodes. Regardless, this increases the scale of shape-based approaches by an order of magnitude or more.

As with the earlier work on hard-sphere representations, a colour forcefield was later introduced to more directly incorporate pharmacophoric information (ROCS colour). As far as the authors are aware, no published work discusses the correlation between the information content of the simple Gaussian and this extension with atom-typing, although it has an observably better performance in most tasks with this feature enabled [64]. In practice, ROCS is either used in shape-only mode, or as a combination mode (ROCS Combination), combining similarity under the shape and colour approaches. However, other researchers have built feature-engineering approaches, essentially descriptors, upon the colour component alone [65, 66]. The authors have been unable to find a canonical reference for the implementation of ROCS colour, although a publication by the developers [64] describes it as "an overlap of groups with like properties (donor, acceptor, hydrophobe, cation, anion, and ring)", which is then added to the shape score. Work by another group [65] states that this score assesses the overlap in 'dummy' atoms possessing that type, and is a simple unweighted sum of the overlap volumes. Interestingly, the authors note that ROCS colour has no internal model for the relative disposition in space of these pharmacophoric features, leading to spurious scores, both positive and negative. The primary issue with alignment-based approaches are their speed: on the orders of thousands of comparisons per seconds, which limits the range of chemical space that can be considered. Attempts to improve this, through GPU-acceleration, have reportedly accomplished speed-ups of 2-3 orders of magnitude [67].

Gaussian approaches - MolShaCS, SHAEP, SABRE, Shapelets

Other approaches to alignment-based similarity comparison have been published, covering similar approaches, such as MolShaCS [68], which features a novel charge-based extension, also utilising the Gaussian product approach. There are a small family of methods, which, although operating on a similar principle, integrate interesting heuristic approaches to render the task of generating a reasonable initial alignment easier, such as triplet-matching methods SHApe-FeaTure Similarity (SHAFTS) [69] and PhaseShape [70]. In a similar vein, Shape-Approach-Based Routines Enhanced (SABRE) [71, 72], makes a sensible initial placement based on shared features, encoded in a reduced graph format, before performing full shape optimisation. A novel variant of the Gaussian-approach [73–75] considers shape as a local property, whilst allowing for global comparisons. The latter approach, 'Shapelets', focuses on capturing shape with high fidelity, rather than virtual screening as such, but still represents an interesting perspective.

Spherical Harmonics (SH) representations capture molecular shape in a unit sphere reference system, where the molecular surface is described at each point by its spherical coordinates and a polynomial radial function. In essence, these harmonics form a basis on the unit sphere in much the same way that familiar trigonometric functions do on a line, or on a circle. This approach was extensively adopted for the comparison of macromolecules [76, 77] and pockets [78], but suffers from the complexity of calculating a reasonable overlap for high-throughput small-molecule screening. Nevertheless, combining SH with pharmacophoric scoring has led to some promising results, with broadly similar performance to the ROCS combined approach on a small test set [79]. To extend the SH concept into a more useful format for Ligand-Based Virtual Screening (LBVS) several attempts at reformulating the underlying description were made, creating "rotation-translation invariant fingerprints" [80], and extracting partially-invariant representations from the distribution of radial terms [81], neither of which is provably invariant under rigid transformation, but which do reduce the complexity of comparison considerably.

1.2.5 Alignment-independent methods

Distribution-based: USR, USRCAT, UFSRAT, ACPC

A simple example of the second major grouping of shape-based approaches, the alignment-independent methods, is the vector underlying simple Principal Moments of Inertia (PMI) [82] plots familiar to medicinal chemists. One branch of this category are the distribution-based methods. Early work in shape-matching led to approaches based on histograms capturing information on distance distributions in atom triplets [83–85], which suffered from a few issues, with initial implementations being limited to molecules with an equal heavy atom count, and later work generating representations requiring several multiples of the memory requirement for coordinate storage to express the distribution adequately. Attempts to remedy these issues led to several useful heuristic approximations for increasing shape search speed, including the calculation of signatures from the calculated vectors, using the properties of hashing functions to provide a good first-pass search of chemical geometric space [86].

Adopting a slightly-different tack with regards the description of molecular shape, but remaining within the domain of alignment-independent methods, Ultrafast Shape Recognition (USR) [87] is a popular moments-based approach to capturing geometric information. It has the advantages of extremely-good computational performance, in terms of memory and time requirements, and a relatively simple conceptual framework. Essentially, for each molecule, four points of reference are defined, two of which are entirely geometric, and two are more tied to the molecular structure. More properly, these are the molecular centroid and the farthest point from same, and the closest atom to the centroid and closest atom to that farthest point. For each of these reference positions, the distributions of Euclidean distances to all atoms are characterised in terms of their first three statistical moments; namely their mean, variance, and skewness, and formed into a 12-element vector for comparison. Slight alterations, to improve the performance, and limit the impact of the second and third moments, were detailed in a patent application and subsequent papers [88, 89], which consist of taking the square- and cube-roots of the second and third moments, respectively, so that the resulting values are expressed in the same unit as the first moment. This also has the effect, given the magnitude of the three moments so normalised, of more highly weighting the first element. This has led to some confusion in the literature surrounding this approach and its implementation [90]. In addition to the speed and memory requirements mentioned above, this approach benefits from a useful website for rapid searching [91] of a subset of the ZINC [92, 93] database, screening millions of compounds per second, an implementation in the popular open-source cheminformatics package RDKit, and a series of small-scale, target-focused prospective applications [88, 94, 95]. The purely shape-based approach mentioned above has been extended and combined with various sources of molecular information, such as MACCS [96], CREDO atom types [97, 98], and information on lipophilicity, electrostatics and chirality [99, 100]. As such, it seems a readily-extensible concept, and a good basis for investigating specific hypotheses.

Another simple approach using a somewhat similar concept is the Auto-Correlation of Partial Charges (ACPC) method, [101], which, as the name suggestions, is based on spatial autocorrelation of point representations of partial charge, and similar in form [102] or intent [103] to earlier work. Although very simple, its performance in a small retrospective study was broadly similar to the other alignment-independent methods so far discussed. The well-established Weighted Holistic Invariant Molecular (WHIM) [104, 105] descriptors could be regarded as members of this class, as they focus on the decomposition of atomic co-ordinates within a defined reference frame.

One recently-published approach, to some extent diverging from established methods, are the Extended Three Dimensional FingerPrint (E3FP) fingerprints [106], which generate the typical atom invariants associated with the ECFP-like fingerprints (*vide supra*), but describe their relations in a Euclidean rather than topological space. They found this method to have a superior performance to the purely 2D ECFP4 in a variety of tasks. Interestingly, they found no additional benefit from the inclusion of explicit stereochemical information; all improvement in task performance was correlated with the inclusion of non-bonded but close atoms. Naturally, deconvoluting the contribution of stereochemistry in this case is difficult, but it seems to have been naturally captured in the formulation of the fingerprint.

Grid-based approaches, which site a molecule within a grid context and calculate some parameters thought likely to describe the activity of the molecule in an attempt to characterise it, have been discussed extensively in the literature. Various approaches, like Comparative Molecular Field Analysis (CoMFA) [107] and its intellectual successors Comparative Molecular Similarity Indices Analysis (CoMSIA) [108] and Comparative

Molecular Moment Analysis (CoMMA) [109], illustrate the benefits and issues of such approaches. The evolution in this approach highlights issues with alignment-dependent methods, and those where interpretability is an issue. CoMMA and GRIND [110], which are based on a similar approach, follow this strategy [111], focusing on developing the method into an alignment-independent approach, for many of the reasons previously discussed. Molecular Interaction Fields (MIF), such as those generated by the GRID [112] program, follow a similar approach on a conceptual basis, and have also been developed into alignment-independent methods adopting the spatial autocorrelation approach [110]. Such approaches have gained wide acceptance in the field of 3D-QSAR, on which there are several excellent reviews [113, 114].

These grid-based approaches typically treat the volume surrounding the molecule in terms of electrostatics, energies, and steric parameters. Another means of defining such a grid would be solvent-based, representing the packed volume for instance. It is possible to define the interactions with this volume through one of a series of surface representations, commonly either the solvent-accessible (Lee-Richards [115, 116]) or solvent-excluded (Connolly [52, 117, 118]) surfaces. One means of conducting shape-based screening using such a representation is to treat such a representation as a collection of curves ([119]) or patches, and use a clique-detection approach ([120]) to identify a matching subset of shape-areas, described by their local curvature, binned into one of five categories in the latter case. This can, of course, be extended with other methods, such as those described in the SURFCOMP [121] approach, which considers near-patch environments as well.

Interestingly, in several instances purely geometric findings, used in various applications (computer game design, industrial computer graphics, engineering) have been applied successfully to the question of shape-comparison, using bounds defined by solvent surfaces, or by van der Waals' radii of constituent atoms. One such approach, shape signatures [122–125], is based on the ray-tracing technique used in the generation of high-quality graphics, provides an elegant formulation for shape as a set of probabilities defining a distribution of segment-lengths for a ray-trace. In essence, this involves allowing a path to propagate into a molecule, reflecting it off an internal surface (where this is constructed from the Lee-Richards surface), noting the distance between the last point of reflection or the origin as appropriate, and allowing this process to continue until a certain number of reflections have been recorded. The descriptor is then formed from a simple histogram representation of the lengths of all segments. This approach has also been extended to incorporate surface properties, to consider shape as a fragment property [125], and to create a variant less sensitive to deformation [126]. This latter approach again calls into question what we mean by shape, and whether deformation-invariance is necessarily something we want in this case. It is relatively common, and rather tempting, to conflate conformation with shape, but as the latter concept is defined on a more or less per-case basis, conformation (atomic position) is not necessarily any more meaningful than surface localisation, or intra-surface distances; it is simply more intuitive for humans.

Another method, based on α-shapes [127, 128], seeks to characterise sets of points using a so-called α-parameter. Intuitively, this is a measure of how closely one fits a 'net' to a set of points, in this case atomic coordinates for a molecule. Expanding on this a little for the 3D case only, for a molecule defined by a set of points of its surface, S, a finite set in R^3 , α is a real number in $0 \leq \alpha \leq \infty$. S is not necessarily convex, nor connected. One could imagine describing the protein surface near a ligand as an example of the latter case. At ∞, the α-shape is equivalent to the convex hull of S, gradually becoming a more detailed representation as α tends towards zero. One could think of α as a parameter describing the radius of a ball forced onto S with a smaller ball more naturally fitting the form of the latter. After this, the representation is faceted for simplicity, resulting in a generalised polytope, with few limitations regarding connectedness or convexity. As such, the approach is related to the mathematical dual of Delaunay triangulation and Voronoi decomposition [128].

Moving on to the applications of this theory, the α-shape Joint-Density (AJD) [129] method represents an elegant implementation of this theoretical concept for a "real-world" problem. Rather than considering the spectrum of α-shapes, they simplify this by taking the 'optimal', i.e. the lowest value of α which produces a single α-surface for which all points are within the described volume. The orientation of each facet is determined by ray-tracing, following which the joint distribution of distances and change in orientation for all pairs of facets is calculated and binned, resulting in a 2D-matrix representation. The authors found that it captured 'shape' well, on a subjective level. As implemented, this approach is a global descriptor of space. An important caveat in their work is that the definition of α-shape employed is technically only valid for homogeneously-sized atoms, in their case defined as carbon. Based on the relationship with the Delaunay triangulation, a size-variant, weighted scheme is possible which should allow for the variation present in actual molecules [130, 131]. One advantage of this approach is in its degree of abstraction: by being less directly related to a specific biological or chemical concept, this framework is readily extensible to assessing the shape of other objects.

1.2.6 Target similarity

Naturally, a method which can be applied to describing small-molecule shape can be extended to larger, macro-molecules. Capturing geometric information about a pocket, or describing its inverse-pharmacophore, allows us to build models describing the relationship between properties of small- and macro-molecules, facilitating target prediction, for instance. It has been suggested that shape-complementarity determines a large part of the ligand-specificity of a pocket [132], with the remainder owing to chemical compatibility and energetics [133, 134]. As several excellent reviews of the topic [135–140] have been published, here, we will focus on shape-based, geometric approaches. Additionally, sequence-based, or sequence-enriched, structural approaches have been attempted.

It has been found that, depending on the definition chosen, approximately a third of a given binding site [132] is typically occupied by ligands, as determined by analysis of crystal structure. Whether this mismatch is owing to a "wasted opportunity", the limits of "drug-likeness", reflects on the dynamics of the interaction, or simply indicates that our definitions of a pocket are rather poor, is unclear. Estimates for average small-molecule binding pocket volume vary significantly, from 610 [141] to 930Å^3 [134], depending on whether an energy or geometry-based approach is utilised. Given this, it is perhaps surprising that most pocket-comparison approaches to date have taken a global, or defined-locality, approach to pocket definition, by which we mean they either compare entire pockets, howsoever defined, or allow a user to specify a pocket region, and treat that as the global segment.

Broadly-speaking, these approaches can be divided into similar tranches as the geometric descriptions of small molecules; alignment-based, and alignment-independent, at the highest level. The former includes methods such as geometric hashing [142–144], and procedures dependent on rigid transformations (roto-transformation, primarily). As implied in our discussion of ROCS and related approaches, the computational cost of this, setting aside the matter of flexibility, is considerable, and so these methods typically depict the local environment in terms of the coordinates and volumes of their Cα atoms. It has been observed that although sequence- and structure-based similarity, when the latter is based on alpha-carbon backbone overlap, are useful [145], they fail to capture the whole gamut of substrate recognition [132]. In general, these binding-site similarity methods take a simplified representation of a known binding pocket, the level of abstraction depending on the algorithm, and then proceed to align and score these representations [133, 146]. Given the profusion of these resources, readers are referred to an extensive review for a complete enumeration [147], to a more focused work for discussion of the algorithms not mentioned here [133], and to a very recent benchmarking study for comparison of outcomes [148].

PocketMatch [149] takes all atoms within 4 Å of a crystallographically solved protein–ligand structure as its basic definition of a binding pocket. To encode the shape of this local environment, amino acids are represented as a combination of the location of their alpha- and beta-carbons, and a measure of the average position of the side-chain atoms. The distances between all pairs of atoms on that local surface are calculated, and the atoms are grouped according to the characteristics of their amino-acid parents, and are then grouped and binned into a matrix representing the interpoint and inter-group distances. When querying these models to look for structurally-related pockets, a simple greedy alignment is performed to assess the similarity of the matrices, and then scored based on the average of the matching elements over all elements in the larger of the two matrices.

CavBase [150] adopts an approach used in generating representations of ligands and inverts it, mapping chemical properties onto the amino acid residues, and then onto the protein surface. From here, it builds up a graph model of how these properties are distributed in space. These graphs are then compared with a clique-detection algorithm [151] to find the maximum overlap between site models. Further optimisation is performed based on aligning the regions of the protein surface associated with each feature.

IsoMIF [152, 153] aims to compare the regions in binding sites in which the binding-critical events take place, using the MIF formalism described previously. With regard to the characteristics of this field, it is parameterised on physicochemical functionalities, some of which are modeled as exponentially decaying relationships. Essentially, the binding cavity is filled with probes, for each of which a molecular interaction-likeness is calculated. As per CavBase, it makes use of the Bron–Kerbosch clique detection algorithm to find the maximum common subgraph when in query mode, which are then scored with a Tanimoto coefficient based on node-matching between the graphs. Using a different fundamental representation, eF-seek [154] considers comparison of the vertex-described Connolly surface as a maximum clique problem.

Surfing the Molecules (SuMo) [155, 156] begins by calculating the local density for each atom in an identified binding site. This measure is the atomic weight of all atoms within a radius, r, of the start atom, over r. In addition, a local center of mass is computed (the center of mass of the aforementioned sphere), and assigned as a vector directed toward the protein interior. A similar procedure is carried out for different "chemical groups," which are, loosely speaking, amino acids grouped by physicochemical nature. For the matching process, both of these sets of triangles, and associated vectors, are used in a simple, rule-based geometric system.

Another purely-geometric approach is PocketPicker [157], which calculates a measure of buriedness from a grid placed over a target molecule, identifying those which are deeply buried within cavities (but not subsurface). Following this, a shape descriptor is calculated, relating the distances and buriedness of the remaining buried grid points, which allows for pocket comparison and prediction of binding site ligandability [158]. As previously alluded to, the geometric SH approach, and derived, roto-transformation invariant representations such as the 3D Zernike descriptors, [159, 160], can also be applied to pocket and protein comparison. The α-shape approach has been utilised extensively for describing and comparing pockets [161, 162], and proteins, with some approaches utilising the weighted Delaunay approach [162, 163]. Of these, CAST [163] is probably best-known. Given the known importance of local properties of pockets [164], relatively few approaches [150, 154, 165–167] that we are aware of have considered it explicitly. Using the α-shape approach, one group defined a local application of the α-shape [168], considering the variation in curvature across a small local area defined by nearby atoms. We are not aware of any subsequent applications of this approach.

An interesting geometric approach to the description of binding pockets (to the best of our knowledge, the authors did not use it to explicitly compare them), is the local roughness indicator (LoRI) model [169], based on the widely-applied mathematical formulation of Fractal Dimensionality (FD). The authors developed a machine-learning model, utilising insights gained from a statistical analysis of the local roughness of binding pockets to predict binding sites in unseen targets. This approach essentially quantifies the complexity and folded-ness of an object, or, to put it another way, how densely-packed the surface of an object is in the volume it occupies [170](for 3D objects). In this case, this is the Connolly surface of a binding pocket (they also made a statistical analysis of ligand FD distributions), resulting in a single numerical representation of shape. Although this value can be approximated with the α-shape approach [171, 172] for 2D objects, this is computationally laborious. However, it does reflect some joint utility in the two approaches. Major advantages of the FD approach are that it is natively scaled, and as such it does not require a weighting scheme, as with the weighted Delaunay triangulation, that it produces a single characteristic number, that it is readily calculable for any arbitrary set of points, meaning that surface locality can be defined however the author sees fit, and, in common with α-shape, that it has a long history of applications and development across multiple fields, demonstrating its utility as a general descriptor. Given the technical advantages of such an approach, and its ready extensibility to other problems, we investigated its utility as a unifying framework for the comparison of small- and macromolecules. As such, we will expand a little on the underlying principles of fractal dimensionality here, to demonstrate its elegance as a foundational concept for shape-based similarity, and, more importantly, to illustrate its utility and ideal properties.

1.2.7 Fractals and fractal dimension

Fractals, as defined by Mandelbrot [173], pre-date their definition. Close approximations of the mathematically exact form can be found throughout nature. We will restate Mandelbrot's initial, intuitive formulation for clarity's sake. If we take as our object the measurement of the length of an island's coastline, using a ruler, we would find it to be a little more involved than initially imagined. Our results would vary widely, having circumnavigated the island, depending on the length of the ruler employed. A ruler of 100 km would give us a rough measure, 1 cm finer, and 1nm finer still, although at this scale, moment-to-moment variation would outweigh the accuracy of our measurement. However, intuitively, as our scale of measurement grows finer, our measured length increases hugely, without a similar expansion in the area of the circumscribed area [174]. In addition, the area of the line itself is still zero. This seeming paradox is at the root of the concept of fractal, or fractional, geometry: How should one describe an object which behaves so "badly"?

This example illustrates the less intuitive side of this branch of mathematics, while introducing an important concept, one of 'characteristic length'. This is both simple, in that the characteristic length of a 1D object is its length, a 2D object its area, and a 3D object its volume, and surprisingly profound. For a line, a square, a cube these concepts have an obvious translation into our understanding of the object. For less well-defined objects, it raises a question, though: what is the characteristic length of an island, or of a tree? That is, how should we characterise natural objects?

Fractal geometry has been extensively applied to facilitate the analysis of such objects, and has proved a useful framework for describing physical phenomena varying from turbulence [175, 176], to the analysis of radiograms [177–179] and homeostatic mechanisms [180], the design of biomaterials [181], antennae [182, 183], novel chemical constructs [184–186], and in the analysis of biological entities [169, 187–192]. As such, we emphasise that this method is simply a formalism, a means of describing the world, which happens to be useful, and to have some properties which render its application appealing.

Previous works have demonstrated the utility of fractal dimensionality as a means of describing and utilising the local roughness of macromolecular surfaces. Early work [189] illustrated that macromolecules are not equally irregular over their entire surfaces, and that there are some general trends relating local roughness and interaction membership. Lewis et al. observed that ligand-binding active sites were smoother, on average, than protein-protein interaction surfaces, and smoother even than the whole-molecule average. Their approach, while considering FD as a local property, delineates this on a patch-wise basis defined by division of a spherical projection, rather than as a property of local atoms. Later work, such as that by Pettit et al. [164], alters the locality definition used, by tying it explicitly to an analysis of the surface surrounding solvent-exposed atoms. This approach facilitated their analysis, in which they showed that surface roughness is a somewhat better indicator of the propensity of a given surface patch to engage in ligand-binding than concavity alone. Interestingly, they found that smaller binding sites are rougher, on average, than their larger cousins, supporting their working hypothesis that a certain degree of geometric complexity is necessary to promote sufficient specificity, and that smaller sites require higher complexity as they present fewer possibilities for non-specific interactions. In contrast to the work of Lewis et al., Pettit et al. found that small-molecule binding sites are rougher than average for their parent molecules, a contrasting result which the authors of the latter study attribute to errors introduced by the earlier spherical projection, and limited data. This latter perspective is supported to a certain extent by work on geometric complementarity in protein-protein interaction analysis [193].

Utilising significant growth in the number and diversity of deposited structures, Todoroff et al. [169] utilised a data-driven approach, facilitating the development of more powerful models incorporating local roughness for binding-site prediction. First, they replicated the findings of Pettit et al. with a significantly larger pool of solved structures, providing further support for the finding that binding sites are rougher, on average, than their parent structures. Next, they extended their analysis to consider pocket atom environments. Interestingly, the authors found that, when considering pocket definition as an inverse problem, using a bound-ligand to delineate interacting atoms rather than a definition based on surface-buriedness, for example, that the former category contains both rough and smooth patches. They proposed that this patchwork facilitates binding, and developed a ‘local roughness model’, the Local Roughness Indicator (LoRI), utilising information on the local distribution of atomic FD values, to predict binding hotspots. This method had considerable success in a large-scale retrospective study. Interestingly, the authors of this study found no obvious, direct relationship between the roughness of each binding partner, and did not posit a mechanistic interpretation for the observed utility of the developed method, in terms of underlying physical interactions. These approaches demonstrate the utility of FD as a means of capturing biologically-relevant information, but have not, to the author's knowledge, been extended to consider both interacting partners. Returning to the question of characterisation, it is useful to first reflect on what we mean by 1D, 2D, 3D, above. Intuitively, this is simple. Here, we provide an intuitive description of the relation between these dimensions, defraying a more rigorous definition until later.

An intuitive description of fractal dimension

One approach to considering dimensionality is to determine a characteristic measurement of each dimension. For example, the characteristic measure of a 1D object is its length, for a 2D object its area, and for a 3D object its volume. If we consider this for a simple object, such as a cube, then we observe that its characteristic measure goes from δ (the length of one side) to δ^2 (its area) to δ^3, its volume. As such, the characteristic measure of a simple object is equal to $\delta^{dimension}$.

As discussed, for simple object this behaviour is straightforward. However, for a class of mathematical oddities, fractals, this relationship does not hold true. Consider the von Koch curve (Figure 1.1, [194]. If we follow the process described, we end with an infinitely-intricate structure, which has several properties of note. It is 'self-similar', in that its structure on the micro-scale resembles its macro-structure. It has 'fine structure', in that it is detailed at all scales. It has a simple method of construction, and a somewhat natural appearance (See Figure 1.2). It is non-differentiable at every point, and does not have a simple formula for construction. One additional characteristic feature of fractals, possessed by this construct, is that its characteristic measure depends on the scale considered.

We will assume, for convenience, that our initial line (iteration n_0) is of length equal to one. The second stage has four segments, each equal to a third of the initial length, giving an overall length of 4/3 (1.33). Our third stage has 16 segments, each equal in length to a ninth of the initial line, giving a total length equal to 16/9 (1.77). As we see, with each iteration, the length of the curve increases to $4/3^n$. As previously discussed, a true fractal is finely, in fact infinitely, detailed. With a generated shape of this sort, every $n < \infty$ is known as a pre-fractal. As such, taking n as ∞, we

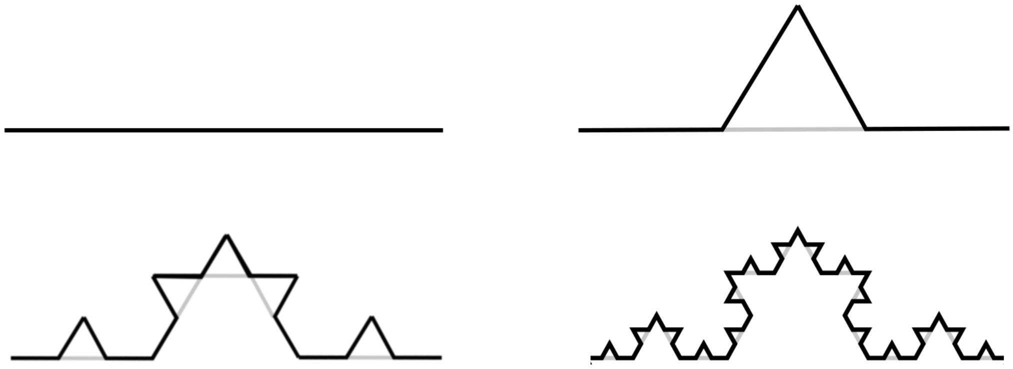


Figure 1.1: Illustration of the initial steps for the construction of a von Koch curve. Taking a simple line, we extract the central third, replacing it with two third-scale copies. We repeat this with each segment. We can continue doing this *ad infinitum*.

find an infinite length for the von Koch curve. However, this infinite length does not alter its area, which is still equal to zero. Intuitively, the length that could be drawn on a 2D object is infinite, as a 1D coordinate provides insufficient bounds. Therefore, length-wise, the von Koch curve behaves like a 2D object, while still being 1D.

This mismatch between the capacity of the familiar dimensions to adequately describe the shape led to the formulation of a novel concept of fractional, or fractal dimensionality (FD). It allows one to capture this expansive behaviour, as with the $(4/3)^n$ growth of the von Koch curve, in a similar manner to the simple relationship described for the cube above. In this case, we have an analytical formulation for the 'fractal dimension' of our curve, where a fractal that is composed of x copied of itself, each at scale $1/\delta$ has an FD of $\log x / \log \delta$. Thus, FD is equal to $\log 4 / \log 3$ (1.26). Relating this back to cube - if we increase the length of the side of a cube times three, we increase its volume by a factor of nine, so 3^3. As such, there are nine copies of the original cube in the later one. Here, we say that such a scaling would result in a characteristic measure which is $3^{1.26}$ times as large.

Of course, in nature, no object is a true fractal. At a minimum, at some stage any pattern will peter out at the atomic level. In practice, it happens much before this. However, this does not render the formalism inapplicable; such objects still obey the weaker definition of a fractal, such that they 'fill space' in a manner more reminiscent of a higher-dimensional structure. Various means of calculating FD for real objects

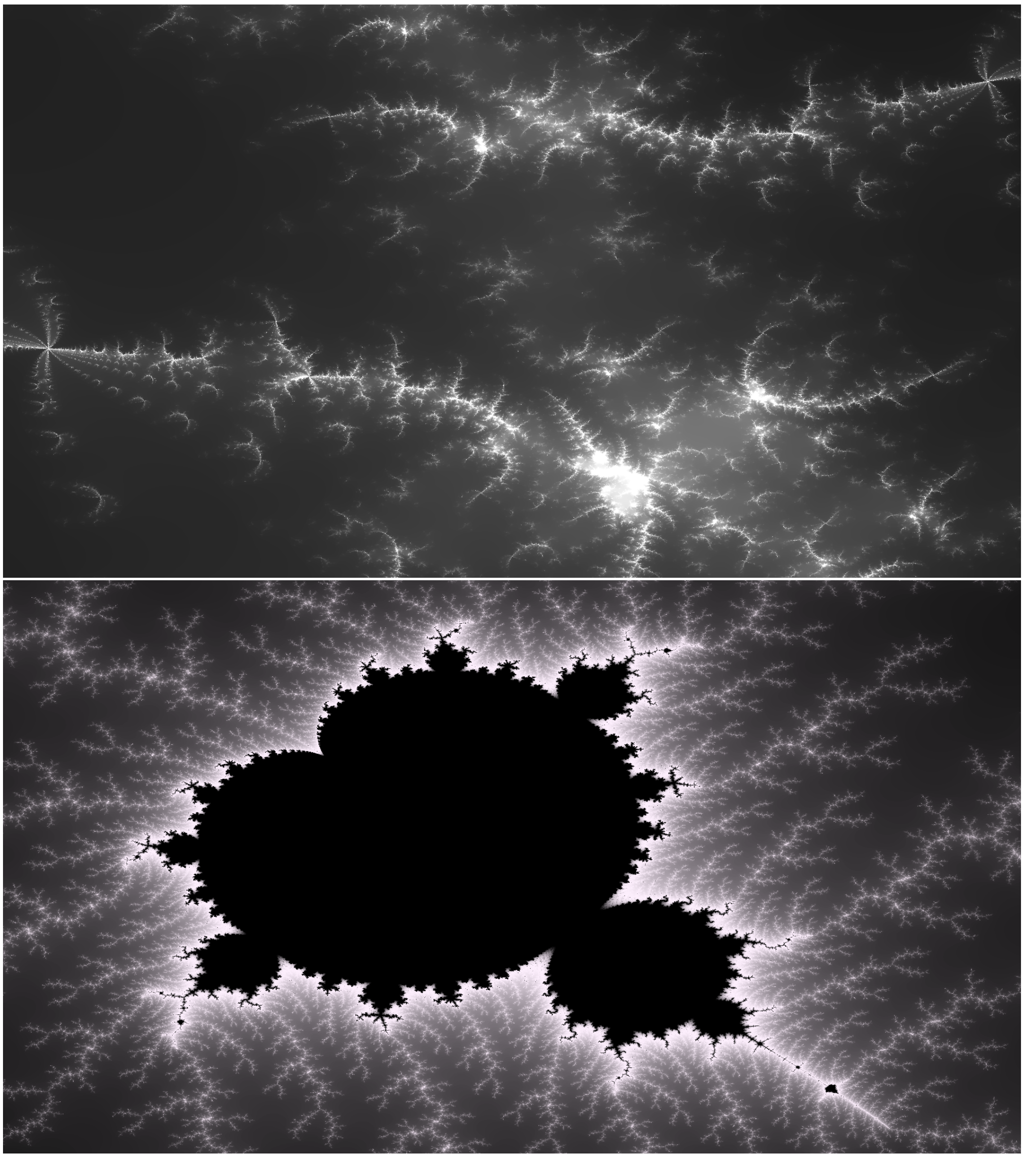


Figure 1.2: Illustration of two exemplar fractals, the ‘explosion’ and Mandelbrot set fractals, each with 4000 iterations of the construction process. Although considerably more complex than the von Koch curve, we see some similarities. In particular, the self-repeating and somewhat ‘natural’ appearance discussed as indicative of a fractal nature are readily apparent. These images were generated with GIMP 2.10.10

exists, which are typically concerned with estimating the relationship between the scale of measurement δ, and the measured size, as with the analytical example above. As we discussed the length of the von Koch curve scales as a power law, accordingly we can take the exponent of a log-log plot of various values for δ, and the measured size of an object at those δ-values, and use this to approximate the FD of real objects.

A formal definition of fractal dimension

Lebesgue Dimension One intuitive definition of dimension, for simple, regular objects, would be that the dimensionality of an object is equal to the maximum number of parameters (or coordinates), needed to uniquely identify a point in the space contained by that object. This does not hold for more exotic constructs, such as a space-filling curve, where the typical formalism used to reflect our intuition is that of the topological (D_T), Lebesgue, or covering dimension. Here, we relist a few mathematical terms for convenience.

A *cover* (C) on S - A collection of subsets on S whose union is equal to S
An *open cover* - a cover where each of its members are open sets.
Open set - a set containing objects up to, but not including, boundary members. In our case, open sets are limited to those with a conventional diameter $< \varepsilon$, for any positive value of ε.
Ply - the smallest number, n, of sets such that each point in S belongs to at most n sets from the cover.

With these terms in mind, the topological dimension of S is equal to the minimum n such that every C of S has an open refinement with a ply of n+1 or less [195]. Usefully, this definition aligns with familiar Euclidean geometry for simple objects in $\mathbb{E}$, such that an n-dimensional Euclidean space, $\mathbb{E}^n$, typically has a covering dimension equal to n.

Hausdorff Dimension Moving on from this, we can describe one of a family of related measures. We will follow the approach of Mandelbrot [196], in stating that:

> 'A fractal is by definition a set for which the
> Hausdorff-Besicovitch dimension strictly exceeds the
> topological dimension.'

The 'Hausdorff dimension', (D_H), introduces a m-dimensional measure in $\mathbb{R}$ such that, when m is equal to n, it is equivalent to the Lebesgue, covering, dimension.

Here, ε is defined as

$$\varepsilon(T) = \sup\{|x - y| : x, y \in \mathrm{T}\}$$

where T is any subset of $\mathbb{R}^n$. Then, let α_m denote the Lebesgue measure of the closed unit ball $\mathbf{B}^m(0,1) \subset \mathbb{R}^m$. For small values of δ, covering S efficiently with countably many sets T_j, with $\varepsilon(T_j) \leq \delta$, and taking the sum of all $\alpha_m \cdot (\varepsilon(T_j)/2)^m$ and the limit as $\delta \to 0$,

$$D_H(\mathrm{S}) = \lim_{\delta \to 0} \inf_{\substack{S \subset \cup \varepsilon(T_j) \\ \varepsilon(T_j) \leq \delta}} \sum \alpha_m \left(\frac{\varepsilon(T_j)}{2} \right)^m \tag{1.5}$$

where the infimum is taken over all countable coverings $\varepsilon(T_j)$ of S whose members have a diameter at most equal to δ. Therefore, as we take the the limit as $\delta \to 0$, the most-restricted infimum cannot further decrease, resulting in the limits $0 \leq D_H(\mathrm{S}) \leq \infty$ (Morgan 2016).

Box-counting Dimension In practice, we typically take a numerical approximation of D_H, such as the box-counting dimension, D_B. This dimension, also known as the Minkowski-Bouligand or capacity dimension, has the advantage of ready comprehensibility and computability. A common definition is as follows: starting with a mesh of cubes of a δ-coordinate mesh on S (where each n-cube is δ^n), let $N'_\delta(S)$ be the number of δ-mesh n-cubes covering S. The box-counting dimension (without qualification) is defined as

$$D_B(S) \leq \lim_{\delta \to 0} \frac{\log N'_\delta(S)}{-\log \delta} \tag{1.6}$$

In practice, we note $N'_\delta(S)$ as $\delta \to 0$, and take the logarithmic rate as D_B. Put another way, we simply take the gradient of the graph of $\log N'_\delta(S)$ vs. $-\log \delta$ as D_B [197]. N.B. Properly, there is an additional term to be considered here, captured by the intercept term in the linear equation, often denoted c, which represents. As we take the limit as $\delta \to 0$, the significance of this term diminishes. For practical cases, where we lack sufficient density to bring δ sufficiently close to 0, our graphical method affords us a simple means to remove it from further consideration.

Correlation Dimension The box-counting dimension, while useful, suffers from a few issues when considering higher-dimensional objects. The difficulty of accounting for variation in starting coordinates increases [198], as does the effect of local sparsity. Essentially, that approach treats all δ-mesh cubes with at least one contained point as equivalent. As a result of the coincidence of these two issues, it has been observed that higher-dimensional objects need substantially increased point-densities to given an accurate estimate of D_H [199]. To mitigate these issues, the correlation dimension (D_C) is intended as a density-sensitive, placement-agnostic estimator of D_H [200, 201].

As with the Hausdorff and box-counting dimensions, we are interested in the scaling of the coverage of a surface with the scale at which it is considered. In this case, our coverage is defined by means of the correlation sum,

$$C(\delta) = \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \theta\left(\delta - \left|\mathbf{X}_i - \mathbf{X}_j\right|\right) \tag{1.7}$$

where $\mathbf{X_i}, \mathbf{X_j} \in S$, δ is the radius of the hypersphere, N is equal to |S|, and θ is the Heaviside function

$$\theta(x) = \begin{cases} 0, & x > \delta \\ 1, & x \leq \delta \end{cases}$$

$C(\delta)$ is an unbiased estimator of the correlation integral. Then, for sufficiently small values of δ,

$$C(\delta) \propto \delta^v$$

where v is a close approximation of D_H, within tight bounds, giving $D_C = v$. The authors note that they were able to get a good approximation of analytical values of D_H with one-thousandth as many points for complex systems.

As discussed, this last approach, the D_C, correlation dimension, has the benefits of stability with low vertex counts, independence from an additional frame of reference other than the point cloud, S, and simplicity. Added to the proven utility of FD for the description of real-world systems, it seems an excellent fit for shape-based description of molecules, and for capturing the complexity of these objects. Henceforth, unless specified otherwise, we shall refer to the D_C of a given object as its FD.

1.3 Molecular similarity comparison

Moving on to our second question, that of how to compare molecular representations once we have them, it is commonly accepted, and with good reason, that "similar" ligands will bind to the same, or similar, targets. Indeed, much of the early, pre-genomic work in protein classification was on the basis of similar biochemical or phenotypic activity profiles in the presence of a molecular probe. Having developed many thousands of molecular representations, we are left with the task of comparing them, and deciding 'how close is close enough'?

In practice, some combination of descriptors and models often proves a more powerful tool than either alone. Fundamentally, the issue is how one should represent a structure which has a complex topological, physicochemical, and geometric nature in a fashion which allows us to compare them meaningfully. Equally, our definition of activity is important: ligand–target affinity has often been viewed as a binary classification (where a molecule has affinity for a given target, or has not) rather than as a regression (molecules have a continuously valued affinity for a target) problem in cheminformatics. This distinction also applies in our concept of ligand similarity itself: while descriptors are often described as being a means of encoding a medicinal chemist's complex, theoretically underpinned, and, to some extent, intuitive view of the chemical world in such a way that a computer can understand it, it is also true that the classification problem (deciding whether two or more molecules are similar to one another, or not) is much easier for trained humans than the ordering or regression problems (e.g., the most similar molecule from a set of 100 is number 47, then 35, etc., or to state that two molecules are 35% similar).

This proves to be a fundamental issue in all cheminformatics endeavours; however, given the comparatively large volume of data a computer can hold in memory at any one point in time, it is not an intractable one, and computational efforts have proven to be increasingly helpful in assisting the efforts of lab-based chemists, biochemists, and molecular biologists.

Similarity Metrics and Indices In the decades since its inception, chemoinformatics has produced many means of comparing such descriptors once calculated, some novel forms of which we will return to in the closing matter of this chapter, which vary in their properties to a considerable degree. In the main, though, a few of these have proved especially useful to date. For the sake of clarity and utility, we will discuss a few here at length. In the first instance, it is important to be clear with the field's terminology. "Similarity," in its applications in chemoinformatics, refers to comparing molecular representations in such a manner that two identical representations would evaluate to one, i.e., 100% similarity.

For these same two representations, their "distance" would be zero. Indeed, distance is the complement of similarity. Obviously, identity comparisons are relatively straightforward, with the nuance being in the non-integer ranges of comparison, where similarity is a non-negative real number. For these methods to perform well, it is necessary to characterise them in a formal manner. Certain characteristics are required to be met before a similarity or distance measure can be properly classed as a metric:

1. Distance values D for two objects A and B must be greater than or equal to zero, and the distance between an object and itself must be equal to zero:

$$\mathrm{A}, \mathrm{B} \geq 0,\ D_{\mathrm{A,A}} = D_{\mathrm{B,B}} = 0 \tag{1.8}$$

2. Distance values must be symmetric, that is, evaluate to the same value regardless of direction:

$$D_{\mathrm{A,B}} = D_{\mathrm{B,A}} \tag{1.9}$$

3. Distance values must obey the triangular inequality:

$$D_{\mathrm{A,B}} \leq D_{\mathrm{A,C}} + D_{\mathrm{C,B}} \tag{1.10}$$

4. Non-identical objects must be separated by a distance greater than zero, i.e., they cannot collide in the metric space:

$$\mathrm{A} \neq \mathrm{B} \Leftrightarrow D_{\mathrm{A,B}} > 0 \tag{1.11}$$

If a given measure meets only the first three characteristic requirements, we call it pseudo-metric; it gives a consistent representation in space of objects, but is not capable of separating all objects in space. Broadly speaking, one might divide the commonly employed metrics into two camps, based on their consideration, or otherwise, of "absent" features in the molecular descriptors [202]. The former camp includes the Minkowski distance metrics (Euclidean, Manhattan, etc.), the latter measures such as the Tanimoto coefficient (Tc) and its complement (for binary-valued feature vectors), the Soergel distance. In the equations to follow, the variables are defined as such: S defines a similarity metric, D a distance metric. For the continuous versions of each function, X_{jA} refers to the j-th element in the descriptor vector for molecule A. For binary-valued variables, a is the number of bits which are 1-valued in molecule A, b the same for molecule B, and c the common on-bits for both molecules. In all cases, formulae are paired in their forms for continuous- and binary-valued feature vectors, respectively.

$$D_{\mathrm{A,B}} = \sqrt{\sum_{j=1}^{j=n} \left| \left(X_{j\mathrm{A}} - X_{j\mathrm{B}} \right)^2 \right|} \tag{1.12}$$

$$D_{\mathrm{A,B}} = \sqrt{a+b-2c} \tag{1.13}$$

Equations 1.12 and 1.13 describe the form of the Euclidean (also L2 norm) distance metric.

$$D_{\mathrm{A,B}} = \sum_{j=1}^{j=n} \left| X_{jA} - X_{j\mathrm{B}} \right| \tag{1.14}$$

$$D_{\mathrm{A,B}} = a+b-2c \tag{1.15}$$

Equations 1.14 and 1.15 describe the form of the Manhattan (also Hamming, city-block, L1-norm) distance metric.

$$S_{\mathrm{A,B}} = \frac{\sum_{j=1}^{j=n} X_{j\mathrm{A}} X_{j\mathrm{B}}}{\sum_{j=1}^{j=n} \left(X_{j\mathrm{A}}\right)^2 + \sum_{j=1}^{j=n} \left(X_{j\mathrm{B}}\right)^2 - \sum_{j=1}^{j=n} X_{j\mathrm{A}} X_{j\mathrm{B}}} \tag{1.16}$$

$$S_{\mathrm{A,B}} = \frac{c}{a+b-c} \tag{1.17}$$

Finally, equations 1.16 and 1.17 describe the form of the Tanimoto similarity coefficients. These metrics vary in their behaviours; some sources [203, 204] state that the Tanimoto coefficient is a better means of assessing the similarity of two molecules, with the other metrics being of more use in placing multiple molecules in context with one another. One disadvantage of the Tanimoto similarity and Soergel distance metrics is their susceptibility to molecular size, owing to a combination of their properties and those of the underlying fingerprint descriptors themselves; the latter tend to be relatively sparse, and are unscaled, so a larger molecule necessarily has a higher probability of exhibiting a given feature, along with other, irrelevant features. This property, coupled with the Tanimoto and Soergel metrics' lack of accounting for shared absent (zero-valued) features, leads to odd size-dependency behaviours. Means of minimising these issues have been discussed in the literature, and mostly involve introducing an additional corrective step for the normalisation of the metric value.

Similarity Threshold Compared to the continued efforts in descriptor design, and in the analysis of their performance in virtual and prospective screening, comparatively little effort has been expended on the matter of determining the relationship between similarity values and similarity in bioactivity. One commonly-used cut-off for Tanimoto similarities was obtained by consideration of local neighbourhoods in terms of their Unity fingerprints, finding that a Tc of 0.85 [205] was sufficient to infer a similar bioactivity, with later work finding that, in fact, this similarity value resulted in two- to three-fifths of the retrieved compounds sharing bioactivity [206]. Indeed, for the popular Daylight fingerprints, it was found that only 30% of compounds retrieved at this similarity level shared activity [207]. As such, it seems that threshold values for a given similarity coefficient likely depend more on the underlying fuzziness of the descriptor, rather than the absolute value of the coefficient, and that even for a given descriptor-coefficient pairing, this might depend on the target domain.

Given the size of typical databases available for virtual screening (the on-demand subset of the popular ZINC15 [92] database has approximately 13 million compounds, as of writing, for instance), one might expect that the nearest-neighbour to an active compound would have reasonable odds of shared bioactivity. This is not necessarily the case, as similarity values have been observed to follow an extreme-value distribution, wherein very few compounds in a given set have a high similarity to a template compound [204]. Thus, the likelihood of finding a "close" neighbour, depends heavily on the novelty of the template compound under consideration, as captured through its descriptor representation [208, 209]. This illustrates issues with choosing compounds based solely on ranking. Even so, if such a compound (Tc $\geq$ 0.85) were found, it would also, likely, be subjectively very similar to the template. This presents another problem: typically, we begin searching in such databases explicitly to locate novel chemical matter with similar bioactivity [210], i.e. “isofunctional but structurally dissimilar molecular entities” [211]. Therefore, in practice, we seek methods which have both a high correlation between their activity values and associated bioactivity, and which retrieve “novel”, that is to say structurally-dissimilar, compounds.

1.4 Benchmarking

As such, given the dynamic tension between our desire to find isofunctional and dissimilar compounds simultaneously, some means of rigorously assessing overall performance is critical. There is, as noted by Nicholls[212], a tendency towards argument-by-anecdote, largely owing, in our opinion, to the difficulty of deconvoluting the various factors determining the performance of a given LBVS run. Briefly, barring innate performance of the descriptor-coefficient combination themselves, we have: the relative “difficulty” of targets or template ligands, the size, diversity, and density of our library with regard our descriptor, and variability in our definitions of activity and novelty [213]. As such, we need both useful metrics, and carefully constructed datasets, to allow us to meaningfully compare our methods. While prospective applications allow for a less biased assessment than their retrospective counterparts in general, they are typically less rigorous, given the smaller dataset sizes.

Methods and Metrics In the main, cheminformatics performance metrics are concerned with enrichment (the relatively early retrieval of active compounds) and diversity (some measure of the intrinsic difference between chemical structures). Here, we take the line of argument proposed by Nicholls, in stating, “in a somewhat circular manner, one of the first characteristics of a good measure is that everyone uses it.” Given that there are relatively few established measures of global enrichment in our field, and as they are observed to correlate reasonably well with one another [214], analysis of the Receiver Operating Characteristic (ROC) by determining the Area Under Curve (AUC) is generally useful. This measure essentially captures the relative ranking of active and inactive compounds under a given ordering, and is closely related to statistical tests, such as the Mann-Whitney U (MWU) [215]. It is unweighted, meaning that it affords no especial credit to the early retrieval of compounds (as do early-enrichment methods), but rather gives a good global assessment of the performance of a method in ordering chemical space. It has a few nice properties, such as being bounded in [0, 1], being broadly independent of the proportion of actives and inactives, and being readily interpretable. A more mathematical definition is provided in Equation 3.7.

Of course, in general, we would prefer not to screen the entire database of compounds. As such, early enrichment methods are popular in the field. Of these, Enrichment Factor (EF), and, more recently, , are the most established. The former, enrichment factor, has the benefits of comprehensibility and a clear relationship with the desired behaviour. Informally, it is simply the enrichment in the observed proportion of actives at some percentage of the ranked compounds, over what would be anticipated were the ranking random. As such, there are two important variables which define the range it can take per experiment: the percentage chosen, and the underlying number of active compounds. As such, it can more properly be viewed as a metric of the performance of the method and the particular dataset chosen, in combination [216]. Additionally, it is insensitive to the ranking of active compounds within the percentage interval chosen. Its bounds depend both on the number of active compounds in the set, and the percentage chosen.

To avoid some of these issues, and to preclude bias by choice of the percentage, the Robust Initial Enhancement (RIE) [217] and Boltzmann-Enhanced Discrimination of ROC (BEDROC) methods were developed. The former avoids a percentage factor by weighting the contribution of each rank position in an exponentially-decreasing fashion. This avoids choosing an arbitrary cutoff, but does still require the user to choose the exponent. In addition, it is not bounded in [0, 1], and depends on the proportion of actives and inactives. Thus, it is not suitable for comparison between cases. BEDROC [216] aims to solve this latter problem, bounding the result in [0, 1]. However, it is still sensitive to the choice of α parameter, which roughly corresponds to the inverse of the percentage term for enrichment factor. Riniker and Landrum [214] have observed that, after accounting for the mismatch between percentage and α terms, that these early enrichment methods tend to be highly correlated. As such, they state that any of these methods, combined with AUC, suffice to describe the early and global enrichment potential of a descriptor-coefficient combination.

The issue of compound diversity is somewhat more problematic. There are two major approaches; one based on neighbourhood analysis, and the other focusing more directly on molecular scaffolds. Neighbourhood analysis and its related approaches depend on the notion of molecular similarity previously discussed to ensure a minimum separation, in terms of maximum or mean values, to a neighbour or neighbourhood of molecules in chemical space [218]. As such, it is highly dependent on the descriptor chosen, and scales somewhat poorly, being impractical for a pairwise comparison of the ZINC screening database previously alluded to, for example, although this is feasible with efficient clustering algorithms [219]. The latter approach, considering molecular scaffolds directly, is somewhat more popular, given its ease of definition and direct relation to the question at hand. TheBemis-Murcko Scaffold (BMS) [26] extracts the backbone of a given small molecule, retaining heteroatom and bond information. This approach differs from the maximum common substructure [220–222] approach by means of generation of an invariant representation which reflects the "decorable" scaffolds of interest to medicinal chemistry, rather than comparison of subgraphs to find commonality. This has the benefit of computational simplicity and an intuitive interpretation. Extending this reductionist approach, Generalised BMS (GMS) [223] removes heteroatom and bond order information, leaving a very "bare-bones" molecular description, but one which captures the general distribution of features. As such, it generates the same result for common bioisosteric replacements, commonly found in series generated through medicinal chemistry efforts. Overall, there is no one "best" measure of diversity, rather it is application dependent. As noted by Kirchmair et al. [224], typically we desire to retrieve novel scaffolds, as these represent a more reliable "jump" in chemical space, with associated claims of biological and intellectual novelty, but these run the risk of interacting with the target in a different manner. This is not necessarily problematic, but does raise the spectre of the "activity cliff" [225], where a rapid diminution in biological activity accompanies a small step in chemical similarity. As such, Kirchmair notes that a reduction in overall enrichment might be acceptable, if accompanied by a concomitant increase in diversity. This will likely remain context-dependent, however.

Datasets and Benchmarking As mentioned previously, many of the metrics vary substantially depending on the composition of the dataset in question. Over the years, benchmarking of LBVS methods has typically been performed on an ad hoc basis, although several datasets have been employed frequently to that end. Recent reviews cover this field in some depth [226, 227], so we will restrict our discussion to a few pertinent examples. The MDL Drug Data Report (MDDR) [228] is a licensable repository of drug pipeline data. As such, it has been employed as a relatively consistent source, manually curated and covering much of the activity of a typical pharmaceutical pipeline. However, given its commercial nature, it is relatively poorly described, resulting in divergent analyses between attempts to utilise it for benchmark studies [219, 229–231]. One criticism is that the sets of active and decoy ligands in the MDDR are often trivially separable based on simple molecular properties [232]. As such, attempting to ascertain the information-richness of a given molecular representation in this environment is somewhat fraught [233].

Datasets specifically constructed for virtual screening have been developed, such as the Directory of Useful Decoys (DUD) [232] and Maximum Unbiased Validation (MUV) [234, 235]. The former, developed for use in docking studies, is constructed choosing pools of decoy molecules which are physically similar to, but topologically distinct from, known active compounds for a pool of 40 targets. As such, and as noted by one of the authors in a later work [236], it is fundamentally unsuited for the comparison of 2D and 3D methods, as it contains many trivial analogues for each active compound. Although popular in the field upon its release [237], and still highly-relevant for docking studies, attempts to utilise it for LBVS approaches led to several criticisms and improvements [238], leading to the later release of the Enhanced DUD (DUD-E) [239] dataset. This considerably expands the number of targets considered, and also addresses issues with the original protocol used for dataset construction, specifically removing trivial analogues in the actives sets and correcting the treatment of charged molecules. However, additionally, they also impose a similarity filter, removing the decoys with a Daylight fingerprint Tc of ≥ 0.5 to the retained actives. The authors reiterate that the enhanced version is also unsuitable for LBVS benchmarking. As such, studies comparing performance of 2D and 3D methods on the DUD set [33, 240] are unlikely to provide generalisable insights.

Moving on to the MUV database, which takes a rather elegant approach to tackling the analogue bias, decoy selection, and data curation issues. They utilise publicly-available data from the PubChem [241] database, subjecting the information extracted from HTS studies to a series of filtering steps, and confirming hits with EC_{50} data. The authors propose some general rules for the construction of an unbiased dataset. First, actives should be embedded within the decoy set, which can be accomplished by comparing distributions of similarity values for available databases. Second, considering simple properties, such as atom counts, the authors state that the distances between actives according to such measures should be at least as large as the distances between the active and decoy sets. Lastly, each target should have actives and decoys with approximately similar dispersion in molecular space. Although these conditions are excellent for demonstrating the ability of a given method to separate active and decoy compounds, it is unclear to what extent performance on this dataset is reflected in the conditions more typical for an LBVS campaign, and, therefore, whether we can use this to gain actionable insights for further screening.

Alternative approaches, which attempt to emulate conditions "in the field", have had some success. REProducible VIrtual Screens DataBase (REPROVIS-DB) [242] collates results from papers with hits from LBVS approaches with $\leq 10\,\mu\text{M}$ potency and scaffolds not present in their template sets. Several reproducibility criteria were imposed, leaving a total of 25 studies which met their defined standard. The authors state that this resource is primarily of use for those who wish to see if their method meets current gold-standards in the field. Alternatively, one could use a database of medicinal chemical findings such as ChEMBL [243] and construct decoy sets according to some similarity criteria [244]. This approach suffers, as with many of the other methods discussed, from poor decoy definition, where uncertainties in declaring a compound inactive affect our confidence in the obtained performance metrics.

Given the diversity of means by which such a set could be constructed, a reasonable approach is simply to try a few of them, and consider the overall and per-dataset behaviours. This method, as published by Riniker and Landrum [214], attempts to set a gold-standard by providing clearly defined subsets, a means of integrating novel descriptors, and a standardised system of metrics and statistics for evaluation. As such, it represents a useful synthesis of the criteria desired for the formulation of a good benchmark [212]. One issue that remains unclear with all such benchmarks, insofar as shape-based methods are concerned, is whether sets based on years of medicinal chemical work primarily developed with a 2D mindset [245] are useful for profiling 3D-focused methodologies.

1.5 Applications

Thus far, we have seen how to define an input space, how to compare locations in that space, and how to assess whether we can meaningfully distinguish between areas of that space containing active and inactive compounds. This leaves us with the final point: what next? What can we do with a well-profiled, well-behaved descriptor, if we can find one?

1.5.1 Prospective screening

Naturally, the primary, and immediately-occurring use would be to look around in chemical space, to see if we can better define the relationship between actives, 'near-actives', and 'never-actives' for targets of interest. Discriminating between the second category and the other two has proven broadly impossible to date, and is a somewhat subjective term at best. As noted by Scior et al. [213], our preoccupation with discovering highly-active compounds, at the cost of promoting structural diversity, is likely a misstep. This approach encourages a certain amount of 'playing it safe', ensuring that methods retrieve known chemotypes at the cost of missing novel ones.

Shape-based methods have a well-established ability to retrieve structurally diverse, active compounds. The importance of building complex screening libraries, incorporating multiple scaffolds, can be captured through analysis of the shape diversity of combinatorial libraries [82]. In that work, the authors used simple PMI plots to describe the diversity of compounds generated through combinatorial substitution of varying numbers of scaffolds, and rationalise the construction of screening libraries with broader applicability to target space than was then commonly the case. As contemporary studies noted [246, 247], the ongoing disappointment with screening simple combinatorial libraries could be correlated with the structural diversity of the targets considered, implying that some more fundamental aspect of chemical structure was missed by such approaches. As such, efforts were placed into the efficient generation of scaffolds themselves, recognising that "compounds that have a common molecular skeleton display chemical information similarly in three-dimensional space" [248], and that increasing molecular complexity in library design improves the likelihood of finding highly-selective compounds [249]. Interestingly, this is also posited as a means of promoting "selective promiscuity", where we tailor a compound to "hit" a group of related targets, based on the shape similarity of their binding pockets [250]. Attempts at clustering chemical space in terms of geometric similarity have identified promiscuous ligands based on shape similarity alone [251]. As well as assisting in the development of diverse, and biologically-relevant, screening libraries, shape-based methods can also help drive the screening process itself [56, 252–254]. For a comprehensive enumeration, we would refer the reader to Kumar and Zhang [90].

The general applicability of an approach is perhaps best illustrated through its utilisation for a variety of targets, and in a variety of scenarios. Various applications, ranging from benchmarking studies to prospective studies in biophysical, biochemical and phenotypic screens [94, 255], have been published. Targets have included protein-protein interactions [48, 256], nuclear receptors [257], and helicases [258], to name a few. As noted by Ripphausen et al. [259], there have been comparatively few large-scale, multi-target attempts to validate shape-based protocols, and, of these, the majority incorporate expert knowledge at some stage to choose compounds, limiting the general utility of the study. Of course, the incorporation of expert knowledge can be useful if the desired end-result is the maximisation of hit rate, but limits the extent to which we can test the shape-based hypothesis. Overall, the increased utility of alignment-based approaches is somewhat uncertain [260]. Whether this is owing to the global alignment metric commonly employed, which struggles with the partial matches often seen in nature [261] is unclear. As Giganti et al. [262] note, the best-scoring alignment is not necessarily the most natural. Their study was primarily focused on molecular docking approaches, but they note that the same effect was seen to a lesser extent with LBVS methods. This might go some way to explaining that there is little observed benefit in including multi-conformer ensembles when screening with the popular ROCS program, for instance [263]. In general, however, the performance of single methods varies substantially depending on the specific use case[255], with our ability to define the domain of applicability of any particular approach remaining rather limited.

Further complicating this picture, shape-based methods are frequently integrated into efforts which combine this basic principle with other approaches, such as 2D or molecular docking, based on a desired outcome. Combination with the former typically indicates an intention to stick to a desired pharmacophore, the latter is often a filtering step to reduce the computational complexity of the molecular docking approach. These approaches have shown some notable successes, and it is likely, in general, that such a combination will be more powerful in terms of retrieval than any one method [231, 264, 265]. To the best of our knowledge, however, no systematic benchmarking of such approaches has been performed with a platform such as that proposed by Riniker and Landrum.

Given the assumed relation between ligand shape similarity and bioactivity, utilisation of such methods for target prediction is a natural extension. Various efforts to that end have been made, with SwissTargetPrediction [266, 267] (based on the ElectroShape [99, 100] USR variant) the most user-friendly of these, providing a web-portal for assessment. Target-fishing approaches, utilising the combined ROCS mode, have also been published [268, 269], which found that incorporation of additional topological information significantly improved prediction accuracy. Given the high degree of structural conservation between binding pockets [270], a different approach is possible. Shin et al. showed that matching of pocket and ligand surfaces was possible to a certain extent by means of the Zernike descriptors previously mentioned [271]. Overall, then, shape-based approaches can be used to enrich and order chemical space, to search that space for compounds whose scaffold had not previously been considered for a given target, and to predict to which (and how many) targets a given compound might bind.

1.5.2 Deep learning in shape-based drug discovery

Given this complementarity, the question arises as to whether it might be possible to generate a ligand de novo to match a given protein binding pocket. To approach this question, we will briefly introduce some terminology and basic material underlying modern progress in the field of machine-learning, and then discuss how it might be used to answer the question of structure-based ligand design.

Parts of this section adapted from:
Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery, ACS Chemical Reviews 2019.
X.Yang, Y.Wang, R. Byrne, Prof. G. Schneider, Prof. S. Yang

For comprehensive coverage of the subject matter, and for insights into other applications of AI in drug discovery, we refer the reader to that work.

AI Architectures

Artificial Neural Networks Artificial Neural Networks (ANN) are composed of interconnected artificial neurons that act as basic information-processing units. [272–275]The product of some input vector $\mathbf{X} = [x_1 \cdots x_n]^T$ and the neuron weights w_i is combined with a bias term b and passed through an activation function $f(z)$ to generate an output. A neuron can be regarded as a function that maps an input vector to an output vector. Mathematically, the output of the neuron can be represented as

$$\mathbf{y} = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{1.18}$$

A typical ANN architecture contains many artificial neurons arranged in a series of layers: the input layer, an output layer, i.e., the top layer, which generates a desired prediction (absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, activity, a vector of fingerprint etc.), and one or more hidden (middle) layer(s) where the intermediate representations of the input data are transformed. A variety of learning techniques can be used to train an artificial neural network, for example, gradient descent algorithms such as stochastic gradient descent (SGD) and adaptive Momentum (Adam), [276] coupled with backpropagation.

Backpropagation refers to the distribution of the calculated error at the output of a model back through its structure, assigning a portion of the total network error to each neuron and enabling the gradient descent algorithm to find a better solution, i.e., the approximation of the underlying function to the problem at hand. Error backpropagation is used primarily for supervised learning tasks, as it requires the target output values for error calculation, although it also works with quantization or construction errors, enabling its employment for architectures such as the autoencoder . For more mathematical details about how the backpropagation algorithm works, we refer readers to several excellent references [273, 277, 278].

The term "deep learning" refers to any learning system composed of several information processing layers [279]. These layers might incorporate multiple machine-learning methods, although this term currently refers mainly to one or more variants of artificial neural networks. A Deep Neural Network (DNN) refers to an ANN that has several hidden layers (with the definition of "deep" being a matter of some debate) with several differences.

1. Unlike traditional ANNs which have typically been used for supervised learning tasks, DNNs can be applied to both supervised and unsupervised tasks.
2. The workhorse algorithm used for training an ANN, i.e., adjusting its parameters, has been SGD coupled with backpropagation. [279] However, this is usually employed in concert with a selection of other algorithms and designs to successfully train a DNN, examples of which include regularisation (such as maxout [280] or dropout [281]), rectification of activation functions, and optimisation of structures, as in Microsoft's ResNet approach [282]. In addition, different optimisers, such as Adam and evolutionary methods, e.g., swarm algorithms, [283] can be employed. The extent to which these improve matters is an area of active research and debate in the AI community.
3. In contrast to using hand-crafted features extracted by extensive preprocessing and feature engineering efforts in traditional ANNs, DNNs can automatically extract useful features from raw input through their hierarchical structure. For example, given a molecular substructure fingerprint representation, for instance, DNNs are capable of learning that a given pattern of bits corresponds to a given feature and that to a certain biological activity. This ability to abstract information, and to generalise, is behind much of the success of these methods.
4. DNN training requires a relatively large number of training examples compared to human learning. Creating a system that can generalise from few examples is the focus of ongoing research. To a certain extent, the availability of data limits the tasks we can accomplish today. ANNs and DNNs are theoretically similar, and therefore, in the following sections, we will discuss them as a single concept. The feed-forward neural network is the most basic form of artificial neural network, where the connections within the network architecture are directed from the input to the hidden layer(s) and onward to the output layer, without loops or backward connections.

Recurrent Neural Networks .

Recurrent Neural Networks (RNN) are designed to identify patterns in sequential data, such as time-series data, genome, and protein sequence data, or simplified molecular input line entry specification (SMILES) strings. In contrast to feed-forward networks, RNNs introduce multiple cells which take as their input not just the current input $x^{\mathrm{t}-1}$ but also the information they have perceived from the previous state $\mathrm{cell}^{\mathrm{t}-1}$, according to the following equation:

$$\mathrm{cell}^{\mathrm{t}} = \sigma\left(Wx^{t} + U\mathrm{cell}^{\mathrm{t}-1}\right) \tag{1.19}$$

where σ is the activation function, W is the weight matrix, and U is the hidden-state-to-hidden-state matrix (i.e., transition matrix). A regular RNN struggles to capture long-term dependencies because of the vanishing gradient problem [284].

Several variations of RNNs were specifically designed to alleviate this problem, such as long short-term memory (LSTM) [285] and gated recurrent unit (GRU) [286]. Deep RNN models based on LSTM or GRU have been used for de novo molecule design. These RNN models [287] are able to capture long-term dependencies and approximate grammars [288], which are necessary to conduct SMILES string prediction since a valid SMILES string, in addition to the correct valence for all atoms, must count ring openings and closures, as well as bracket sequences with several bracket types. All of these elements combine to form a grammar relating chemical vocabulary to molecular features and should be captured by the model to compute valid SMILES that correctly represent the intended molecule.

Autoencoder and its Variants A minimalist autoencoder consists of an encoder functionality, translating an input into a latent space, and a decoder, translating this internal latent-space representation back to the original input space. Given an input $\boldsymbol{x}$, the goal of an autoencoder is to compute a reconstruction $\boldsymbol{x}'$ with minimal error compared to the original input $\boldsymbol{x}$, while also having an internal representation with fewer features than are present in the input. In essence, then, an autoencoder can be regarded as a dimensionality reduction approach, as it can be used to build potentially more meaningful, or less noisy, representations of feature vectors. It is also extensively employed in generative modelling, as it allows for the creation of a well-described and consistent internal representation with few dimensions that can be translated back to the original or a related space.

When the input data have a complicated internal structure, as is often the case with biological data, multiple autoencoders can be stacked, adding nonlinearity and flexibility. Variations of autoencoders have been developed to prevent "vanilla" autoencoders from simply approximating the identity operator and to increase their ability to extract useful features from data. Examples include the denoising autoencoder [289], the Variational Autoencoder (VAE) [290, 291], and the Adversarial Autoencoder (AAE) [292].

Denoising autoencoders are intended to recover the original undistorted input data from corrupted input data, making the final model more robust to noisy data than vanila autoencoders. To accomplish this, stochastic noise is added to $\boldsymbol{x}$, giving $\boldsymbol{x}^*$, and the autoencoder is trained as before, with the exception that we are now interested in the loss values obtained from $\boldsymbol{x}$ and $\boldsymbol{x}'$, when trained on $\boldsymbol{x}^*$.

In the case of a VAE [290], we direct the system to produce outputs similar to our inputs by adding a requirement that the distribution of the variables in the latent space should follow some distribution, most commonly a Gaussian. Subsequently, these latent vectors are fed into the decoder to reconstruct the input. Due to this, we have the possibility of generating new samples that are similar to the original samples used for training, which has the effect of increasing the localised density in our latent space. This process has the advantage that input data that are somewhat similar to our training set are more likely to have a sensible decoded form than in simpler autoencoder models. Besides the reconstruction loss used in vanilla autoencoders, another term is also added to the loss function in VAEs, that is, the Kullback-Leibler divergence between the distribution created by the encoder and the prior distribution.

Since the integral of the Kullback-Leibler divergence term in VAEs adds to the computational cost, AAEs [292] have been introduced to avoid the necessity of evaluating that term. This is achieved through introducing a new network (called the discriminator) to each AAE, which is trained to determine whether its input comes from encoder-generated latent vectors or from a prior distribution determined by the user. The first step of training the overall network is the same as for the vanilla autoencoder above, where we minimise our reconstruction loss, i.e., we train the encoder-decoder system to reproduce inputs. Then, we train the discriminator to distinguish between noise and true values and predicted true values. Finally, we optimise our encoder by using the discriminator with a cross-entropy loss function; essentially, we use the discriminator to label our input data and drive the optimiser to produce only outputs that are indistinguishable from true data.

Convolutional Neural Networks Convolutional Neural Networks (CNN) are used in situations where input data can be represented as images or image-like objects [293, 294]. A typical CNN contains at least three components: convolutional, pooling, and densely connected layers. A convolutional layer is described in terms of its width, height, and depth; that is, it captures x-and y-coordinate information over a small receptive field (a square of pixels), with a depth z corresponding to different sources of information (e.g., RGB colors in images) and uses the patterns observed as it slides across the input image to set weights. The main advantages of the convolutional layers are that they reduce the number of parameters via their weight-sharing mechanism and gradually build up spatial and configural invariance [294].

Pooling layers essentially implement subsampling to reduce the impact of noise and the number of learned parameters. In addition, pooling layers add a certain level of resilience to small shifts in the placement and orientation of input features. A convolutional layer followed by a pooling layer can form a convolutional module, and each module of the CNN network learns to identify features while preserving their spatial relationships. These properties lead to a major advantage of CNNs over standard ANNs, namely, that they are translation-invariant, i.e., they can recognise the same feature in different areas of the input field. Generally, a CNN is composed of a stack of convolutional modules to achieve feature extraction, followed by one or more fully connected layer(s) for prediction and loss-minimisation. It should be noted here that the convolution operation is linear, since the output of each neuron $\varnothing$ in the feature map is simply the result of multiplying the input value x_i by the weights w_i of a given filter and adding them:

$$\varnothing = \sum_i x_i w_i \tag{1.20}$$

Therefore, in most cases, the output of a convolution layer will be passed through some form of nonlinear activation function (such as rectified linear unit, ReLU) to allow the network to handle more intricate relationships. As CNNs grow ever deeper, a new problem arises: information about the input or gradient can vanish and "wash out" when it reaches the end of the network.

Several different approaches have been proposed to address this and related problems, such as creating short paths from early layers to later layers [281, 295, 296] or connecting each layer to every other layer in a feed-forward fashion to ensure maximum information flow between layers in the network [297]. Similar to other neural network architectures, hyperparameter tuning has a great impact on the performance of CNN models, and the number of possible choices makes the design space of these architectures large, rendering an exhaustive manual search infeasible. In this case, various approaches have been developed to provide reasonable initial hyperparameter values and to rationalise the tuning process, examples of which include grid search, random search [298], Bayesian optimisation [299, 300], and evolutionary methods [301]. CNNs were originally developed for two-dimensional image recognition.

Some of the approaches utilised in typical CNN networks, such as the pooling algorithm, are adopted to reduce the dimensionality of the representation and to allow for position shift, which can lead to a loss of information and hence a poor performance in drug discovery-related studies. An advanced deep learning architecture, the Capsule Network [302], allows the modelling of hierarchical relationships of the network's internal knowledge representation and could have considerable potential for drug discovery-related research. CNN architectures can also trivially be applied to sequence transduction problems [303], such as with the ByteNet [304] architecture, which achieves good performance on standards such as English to German translation, runs lightly and in linear time, but with the issue that path-lengths (how far the model has to "look" at any point) can be logarithmic, reducing the working memory.

Attention and Attention-Based Architectures Rather than being a network architecture as such, attention is a means of improving the performance of other models, with special attention to the generative RNN and encoder-decoder architectures, and allowing them to generate more useful output in cases where there are substantial long-range dependencies in the input or output sentences. English-to German translation is often given as an example of such a problem, given that German is a subject-object-verb order language. This means that decoders can struggle to produce sensible input, as they are not sure of the relationship between subject and object. Bahdanau et al. [305] developed an approach to get around a "bottleneck" introduced by encoder compression, forcing a fixed number of dimensions onto the latent space representation of input sentences and avoiding the direction-dependence of earlier methods [306], significantly increasing performance on translation tasks. In essence, the attention model allows the decoder in a bidirectional-LSTM encoder/LSTM decoder model to focus on sections of the input when producing an input, thus "paying attention", rather than using only information in the context vector. Luong et al. [307] extended this work, adding new kinds of attention models, and the distinction between global and local attention. It can be defined as a means of "mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key." [308]. As such, and in tune with our intuitive understanding of the term, attention facilitates 'focus', allowing the model to focus on only the most relevant information when making its next step.

Having observed the power of this concept, a novel architecture, the "transformer" [308] was proposed, which aims to solve some known issues with sequence transduction models implemented with RNN and CNN architectures. As there is no recurrence in this architecture, word or character position is maintained with use of a positional encoding, in particular a sine function. Describing the overall architecture, it inherits its basic formulation from earlier work [309, 310], wherein stacked RNNS are used as encoders and decoders (*vide supra*). Each of these layers contains two substituents: the "multi-head self-attention" and fully-connected FFN. In addition it features some "short-circuit" or highway connections [311], which essentially bypass sections of the architecture, "carrying along" the gradient, to deal with the vanishing gradient issue.

Several kinds of attention are employed here. Multi-head attention describes multiple learned projections of the query, key, and value matrices into lower-dimensional spaces (which are then re-projected), reducing the impact of averaging over multiple attention-weighted positions, whilst still running with the same performance overall as the approach described by Bahdanau [305]. The implementation of self-attention differs between the encoder and decoder stacks. For the encoder, the current state can attend upon all states from the previous layer. For the decoder, inputs to the right of the current position are masked, so the model is unable to "look forward". For a fuller description, the original paper is strongly recommended, also as an introduction to attention-based methods.

General considerations

Constructing a Model Applying a particular AI algorithm in drug discovery is a sequential process that requires the proper definition of the problem domain [312]. This process typically includes problem definition, data preparation, design of the AI architecture, model training and evaluation, and understanding and explaining the results. More specifically, one should be clear about the problem at hand before any specific architectural decisions are made, since the choice of the machine learning method should be appropriate for the problem under study. First, one needs to know whether this particular problem belongs to the domain of discriminative or generative tasks. With the task of AI modeling in mind, the next step is to design an appropriate model architecture. This step includes the choice of a suitable algorithm and setting sensible initial values for hyperparameters.

In general, ANNs are the go-to part of the field for these tasks, given their ability to generalise and hypothetical ability to approximate any input-output relationship function. Hyperparameters vary with different AI algorithms. The architectural parameters for a neural network include, but are not limited to, the choice of the number of neurons and layers (and their type), learning rate and decay, regularisation parameters, and the presence of connections between neurons or adjacent layers. After a provisional architecture is determined, it is time to prepare the data set. The representativeness, quality, and quantity of initial data have a crucial impact on the quality of an AI model.

Once the initial architecture and the data sets have been established, one may proceed to model training and evaluation. The training step aims to search a set of parameters with the objective of reducing/minimising the prediction error. The final AI model should have the ability to express the underlying relationship between the molecular representations and practitioners' own specific purposes. If this is not the case then examining specific examples can help guide the practitioner in developing their model "ecosystem" to accomplish the goal.

Input Data Preparation Newcomers to the application of AI methods in drug discovery projects tend to make improvements on the overall performance of the model by focusing on the deployment of the latest AI approaches. However, it is often more beneficial to focus in the first instance on the training data, as it underpins all further progress. More high-quality data usually leads to a better generalization performance, regardless of model chosen [313]. However, data preparation is a labour-intensive and challenging task. One needs to understand the origin and meaning of the training data, for example, the types and complexity of entities represented, the quantity of data, and more domain-specifically, their distribution in chemical space, for instance. Overall, the question is how well we have populated the space of possible inputs that we might want to make predictions for. If the need for more data is apparent, one must decide on a preprocessing strategy, whether unlabelled data might suffice, what sort of representation would be most useful for encoding the entities represented, etc. Importantly, no single rule is universally applicable. Here, we attempt to provide some guidance on these issues.

Data Types in Medicinal Chemistry The most frequently used input data type in drug discovery is a fixed-length input vector (e.g., molecular descriptors, fingerprints) [314, 315]. However, there are two major limitations inherent to this type of representation. Such vectors tend to be rather large to encode all possible substructures without collisions (overlap), resulting in models that have many learnable parameters and that attempt to learn from relatively sparse inputs. For example, a fingerprint vector of size 43,000 was used by Unterthiner et al. [316]. Similarity assessment in such a high-dimensional chemical space is error-prone [317].To partially alleviate this problem, various types of graph fingerprints [318, 319] have been proposed, which are calculated with a differentiable neural network whose inputs are molecular structure graphs. The other limitation is caused by the difficulty in establishing bijection (one-to-one correspondence) between input vectors and molecular structures. Input vectors can be easily generated from a molecular structure, but reverse structure reconstruction from vectors is an extremely difficult task, especially as a single fingerprint representation likely corresponds to multiple possible chemical structures. One way to avoid this limitation is to use AI deep generative models in combination with SMILES strings as molecular representations [320].Such combinations have been widely investigated in recent years for the generation of novel compounds with desired properties. In this case, the output data is either a SMILES string or a fingerprint vector. The output data in most other AI-assisted drug discovery projects are numerical value(s), with binary values corresponding to binary classification, integer values for the multiclass classification (or clustering), and real-valued numbers involved in regression tasks, often with experimental biochemical data.

1.5.3 Applications in generation and regeneration

Note: This section is shortened and adapted for clarity, please refer to the original work for a comprehensive discussion of the state of the field as of time of writing.

Drug design aims to generate chemical species that meet specific criteria, including efficacy against pharmacological target(s), a reasonable safety profile, suitable chemical and biological properties, sufficient novelty to ensure intellectual property rights for commercial success, etc. With the aid of novel algorithms to carry out the design and evaluation of molecules in silico, de novo drug design is increasingly considered an effective means to reduce the vastness of chemical space into something more manageable for the identification of tool compounds for chemogenomic research and for use as starting points for hit-to-lead optimisation. In this section, we highlight the achievements of AI-assisted de novo drug design and point toward potential future developments. The reader is referred to some comprehensive resources regarding de novo molecular design for further information. [321–323].

Early de novo drug design approaches [322] almost exclusively used structure-based methods to grow ligands within the constraints (steric and electronic) of a binding pocket for the target of interest, whether adapted directly from protein structures or inferred from properties of known ligands [324, 325]. A limitation of these early methods was that the generated structures were prone to synthetic infeasibility and poor druglikeness (such as poor drug metabolism and pharmacokinetic properties). More recently, the ligand-based de novo design method has demonstrated its applicability in medicinal chemistry. Generated compound libraries may be additionally analyzed with the aid of a scoring function which takes into account several properties such as biological activity, synthetic accessibility, metabolism, and pharmacokinetic properties [326, 327]. One way to build such a virtual library is to use a curated subset of chemical reactions, along with a group of available chemical building blocks, leading to a pool of synthetically accessible molecules [328]. This approach was adopted by Hartenfeller et al. [327, 329] in the development of DOGS, a software which allows the 'in silico assembly of molecules' based on a template structure and the aforementioned building block and reaction libraries. Syntheses of these products resulted in novel active compounds against γ-secretase, histamine-4 receptor, and polo-like kinase 1 and in the successful imitation of the pharmacological profile of a natural product, (-)-Englerin A, among other applications [330].

A related approach is to apply knowledge-based expert rules from medicinal chemists to design analogues for a query structure. For example, Besnard et al. [331] used a knowledge-based approach to automatically generate novel dopamine receptor type 2 (DRD2) modulators with specific polypharmacological profiles and suitable ADMET properties for blood-brain barrier penetration. Even though the use of either knowledge-based or reaction rules can reliably and effectively generate novel molecular structures, these approaches are limited by the inherent rigidity imposed by the predetermined rules and reactions. Whether this fact poses a practical problem may be a matter of discussion. For the purpose of scaffold hopping, however, the cardinalities of the virtual compound libraries that can be generated by a rule-based system easily exceed 10^{30} drug-like molecules.

A third approach, called "inverse QSAR", deals with the de novo design task from a different angle. Instead of first generating a virtual chemical library and then scoring and ranking it based on similarity to a template compound, inverse QSAR attempts to find an explicit inverse mapping $y \to X$ from properties y to molecular descriptor space X and then maps back from the favorable region in descriptor space X to the corresponding molecules [332–338]. The major obstacle of inverse QSAR approaches lies in the selection of a molecular representation which is informative and suitable not only for sufficiently handling the forward QSAR task for a given biological property but also for the subsequent reconstruction stage to be meaningful. Many de novo drug design methods utilise sets of molecular building blocks or fragments of synthesized compounds for molecule assembly to reduce the risk of generating unfavourable chemical structures [333, 339–341]. To avoid overlooking attractive candidate molecules and to increase the diversity and novelty of generated structures, a large fragment library should be used. This comes at the price of a substantial increase in the cost of the fragment swapping and similarity search processes. Ikebata et al. [332] investigated the use of a fragment-free strategy for the generation of novel molecules with desired properties by integrating forward and backward QSAR predictions with the aid of machine learning techniques. They first set up a group of machine learning QSAR models for the prediction of various properties of a given molecule. These forward models were then inverted through Bayes' law, and the resulting posterior distributions were used to identify high-probability regions for molecules with desired properties. A chemical language model based on SMILES was created to circumvent the problem of chemically unfavourable structures.

As noted by Gómez-Bombarelli et al., [319] many methods in de novo chemical design depend on explicit rules based on chemical knowledge for replacing or adding molecular fragments to yield new structures, which may bias the search space and ignore certain other structures. Efforts to resolve this issue have primarily focused on means by which to learn such transformations implicitly via generative models. [342]Deep generative networks have shown promise in de novo drug design, without any explicit prior chemical knowledge. Examples of this approach include AAEs [343, 344], VAEs [319], and RNNs [345–347]. Among the various neural networks, there is a growing interest in RNN-based generative models for the de novo design of molecules, [347, 348] given their ability to cope with sequential data with long-range dependencies, such as the SMILES chemical representation format and to learn complex grammar. Harel and Radinsky [349] proposed a prototype-driven diversity network, a generative chemistry architecture which combines encoder, VAE, CNN, and RNN components to generate diverse molecules with similar properties to those of a molecular template. They found that the proportion of valid SMILES, i.e., those parsable into sensible molecular structures, from generated suggestions was significantly improved by prototype-conditioning the VAE. This suggests that not all areas of the latent space representation conditioned by the encoder are equally easy to translate into real molecules, likely because of the sparsity in the training data and insufficient penalization of such events during training. Interestingly, 0.01% of the molecules generated using 869 FDA-approved compounds as prototypes were, themselves, FDA-approved.

Gómez-Bombarelli et al. [319] proposed a generative model which adopts a somewhat similar strategy. The interconversion between SMILES and continuous latent-space representations was achieved bu combining a VAE with an RNN encoder and decoder. To enable molecular design, an additional multilayer perceptron was trained to predict properties of interest based on the latent space coordinates of molecules. The prediction task was jointly trained on the reconstruction task, so that when given the latent vector of an encoded molecule, new candidate vectors can be generated and decoded into corresponding molecules by moving in the direction most likely to improve the target properties. The results showed that the model exhibited good predictive power for electronic properties (i.e., orbital energies). Moreover, upon adopting an optimisation objective incorporating a mixture of drug-likeness and synthetic accessibility, the model was able to perform iterative, gradient-based optimisation to suggest molecules better matching the desired properties.

Segler et al. [347] demonstrated that RNNs trained on the SMILES representations of molecules can both learn the grammar required to generate valid SMILES and generate candidate molecules with similar properties to those of template compounds but with differing scaffolds. The de novo drug design cycle of this method adopts transfer learning, in which an RNN model is first trained on a large set of molecules and then further retrained with a small set of active molecules to bias the sampled molecules toward a given template set. Their retrospective results showed that their de novo RNN model could reproduce 28% of 1240 known active compounds against *Plasmodium falciparum*, without having seen the compounds in the initial training, having utilised a roughly equivalent number for fine-tuning. For *Staphylococcus aureus*, the corresponding figures were 14% of 6051 test molecules, having trained on 1000.

Similar to the work of Segler et al., Yuan et al. [350] described a new library generation method, Machine-based Identification of Molecules Inside Characterized Space (MIMICS). This method consists of two steps. The first step is to use a character-level RNN (char-RNN) to learn the probability distribution of characters in SMILES strings for given chemical subsets, followed by postprocessing to eliminate structures with invalid valences, aromaticity, or ring-strain issues, resulting in MIMICS output of molecules with similar properties and dissimilar scaffolds to those of the input set. MIMICS-generated compounds were found to act as inhibitors of the unfolded protein response (UPR) and VEGFR2 pathways in cell-based assays, demonstrating the capability of MIMICS to generate useful, novel compounds. The ability of RNN-based generative approaches to suggest molecules with similar biochemical activities to those of a template or set of templates but with novel scaffolds has been the subject of much interest.

For example, Merk et al.[351] developed an LSTM model to generate novel bioactive mimetics of natural products with retinoid X receptor modulating activities. It adopted a transfer learning approach, learning the basic grammar of small molecules from ChEMBL and was then fine-tuned on a small set (N = 6) of known natural product retinoid X activators. The generated compounds had a distribution of natural product-likeness 483 scores intermediate between those of ChEMBL and the dictionary of natural products, and 50% of the synthesized molecules showed activity against retinoid X receptor. This approach allows the generation of compounds which incorporate natural-product-like features, while still retaining some of the synthetic feasibility of typical small-molecule compounds and promoting structural diversity.

Arús-Pous et al. [352] demonstrated that two-thirds of the GDB-13 chemical space, constructed by means of a rules-based enumeration scheme, can be efficiently reconstructed with an RNN generative model trained on less than one percent of the input space. They found that the model struggled to reproduce more complex structures but performed well overall. The well-defined subset of chemical spaces allowed a rigorous comparison of the model performance to that of an ideal generator (one which only produces valid SMILES, all of which are part of GDB-13) by using the coupon collector problem to establish the baseline performance. The proportion of valid SMILES in the output of any given generative model is a commonly adopted metric by which to evaluate reconstruction performance.

One major contributor to the generation of invalid SMILES is the long-range dependency issue, wherein, for example, the opening and closing brackets representing a ring structure might be separated by many intermediate characters, resulting in an increased likelihood of unclosed rings in the output SMILES. Pogány et al. [353] addressed this issue through use of a bidirectional LSTM architecture coupled with Luong global attention, [307] a strategy which has been observed to improve the performance in terms of long-range dependency sequence generation tasks in other fields. In addition, their approach used reduced-graph representations as intermediate descriptions of molecules and then employed these representations to generate corresponding SMILES which met the pharmacophoric template. Lim et al.[354] proposed a molecular generative model incorporating the conditional variational autoencoder for de novo molecular design. This approach concatenates molecular property information to the latent representation of molecules. The performance of the model was demonstrated by generating drug-like molecules with specific values for five target properties (molecular weight, partition coefficient, number of hydrogen bond donors and acceptors, and topological polar surface area) with a defined margin-of-error and by creating analogues with variable log P values while constraining the other properties.

An approach to this issue is to adopt the transfer learning method, which aims to improve predictive performance by using insights gained from training on a previous task and transferring them to a new but related task. Awale et al. [355] trained LSTM generative neural networks using molecules taken from commercial catalogs and from FDB-17 499 (a database of fragments up to 17 atoms) and performed transfer learning with ten drug compounds to generate new analogues of these drugs. Their results suggested that transfer learning can learn the rules to assemble small fragments into larger, drug-like molecules and that the performance was broadly similar regardless of whether the models were trained on fragments or larger molecules. Gupta et al. [356] successfully integrated the transfer-learning approach into RNN-based generative models via a new approach to de novo drug design. An LSTM-based RNN model was first trained on ChEMBL to learn the correct SMILES grammar. The transfer learning technique was used to fine-tune the model to produce SMILES strings which were structurally similar to small libraries of target-focused compounds. Starting from a single receptor-binding fragment, the researchers demonstrated that their generative RNN model could successively grow the remaining molecules. They found that even with a small number of representative molecules used during the fine-tuning process, their approach generated structures with similar chemical characteristics to those of the provided subset providing a means by which to carry out hit-to-lead optimisation with limited data.

Blaschke et al. [320] compared the performance of adversarial autoencoders as structure generators with several VAE instances. The VAE instances were named teacher and no-teacher VAE, with "teacher" referring to teacher-forcing, a technique in which a model gives both the output of the model and the corresponding character from the training corpus as inputs for the next time-step. Molecular structures were encoded into a continuous latent space and then decoded into the original space to determine the loss introduced by this compression process. The reconstruction accuracy of all models was at least 95%, with an observed benefit of the adoption of teacher-forcing in the VAE case during training, with the effect inverted during generation. The overall reconstruction accuracy was higher for all AAE models considered and best of all for the uniform AAE, which forces a uniform distribution onto the latent vectors. If the latent space is well constructed then the distance between compounds in their original space should be preserved. This behavior was used to sample latent space vectors at increasing distances from the vector corresponding to celecoxib, leading to analogous compounds being proposed, and to confirmation that the distance in latent space corresponds with the Soergel distance between the compounds' ECFP6 representations.

An interesting application using the same architecture, the VAE, was recently published [357]. In contrast to much of the existing literature, it incorporated information about molecular shapes and pharmacophores, as encoded through a 3D-CNN, using an LSTM model to caption the result, and, in doing so, generating a SMILES representation. Attempts to regenerate test SMILES were minimally successful (1.74% of cases), and suffered from a high invalid rate (65% were parsable after disabling of valency and sanitisation checks), with the primary issue being improper ring closure. For those which could be parsed successfully, aromatic rings were generally present, but hydrogen-bond acceptors and donors were rarely reconstructed. The authors note that the decoded shape representation is not roto-translation invariant. This detailed analysis of reconstruction error, and definition of applicability domain, is often missing from other works. In an additional analysis, the authors compared their results with those generated from some of the other methods discussed here, and found that the molecules generated from the shape approach had a similar distribution of simple properties, and higher diversity than a SMILES-focused VAE model, with a lower shape-similarity (as measured under USRCAT and its associated metric) to the template molecules. As a result, although encoding geometric information does seem to be associated with an increased diversity, it remains unclear to what extent this is owing to noise other than that deliberately introduced via the VAE approach. This approach has the major advantage of being readily extensible to the inverse pharmacophore problem, should the outstanding issues be resolved.

To achieve further acceptance of compounds generated from AI techniques, it will be important to understand the current limitations of this design approach. One of the main limitations is the lack of informative and suitable ways to translate the domain knowledge learned by the AI model into molecular structures. Even though there are various choices for molecular representation, such as SMILES strings, molecular fingerprints, descriptors, or novel representations based on chemical graphs, the most popular method to-date in AI-assisted de novo drug design is the SMILES representation of two-dimensional molecular graphs because of the ease of conversion to a molecular structure. There is still room for improvement of these outputs in terms of validity and chemical novelty.

Nearly all generative models in the field of AI-based de novo drug design barely consider the structural (steric and electronic) information on the target protein, although examples of deep learning networks for predicting protein-ligand binding affinity that incorporate protein structural information exist [358]. The lack of standardised approaches for the analysis and scoring of generated molecules renders it difficult to create an honest appraisal of the merits and pitfalls of each approach. Proxy measures, such as the proportions of valid and unique SMILES, give an overview of the generative capacity but little insight into the representative aspect. Next-generation small-molecule drugs will be designed to interact with multiple targets [359, 360]. This multidimensional view of chemical and biological ensembles is a perfect base for AI-based de novo drug design, as the incorporation of multiple objectives can improve the power of such system markedly.

Chapter 2

Aims

Shape-complementarity is commonly regarded as a powerful determinant of ligand-target binding, and of the promiscuity of both partners. Several shape-based virtual screening approaches, each with their own definition of what constitutes molecular shape, have been widely adopted in academic and industrial environments. Their popularity owes much to the perception that they can aid in ‘scaffold-hopping’. These differ substantially in their complexity, utility, and flexibility.

Existing work profiling their relative advantages and disadvantages has largely been conducted on an ad hoc basis, with the addition of extra ranking and refining methods rendering it difficult to accurately profile each approach. Although some promising work has considered the opposite relation, such that the shape of target pockets defines the shape of binding partners, these tend to be somewhat anecdotal, localised to a particular target family.

An approach which would allow us to retrieve novel compounds for a given target, to describe that target, and, ideally, to suggest ligands based on structural analysis, would help both to assess the importance of shape, and offer a valuable contribution to the field of virtual screening.

Hypothesis: Shape complementarity determines a considerable portion of the target-complementarity of ligands, and the ligand-complementarity of targets, in a manner substantively different to simple descriptors.

To assess the validity of this hypothesis, we followed a five-fold strategy:

1. We developed a method based on the formalism of fractal dimension, allowing for the construction of local and global descriptors of ligand and target shape.

2. We constructed a benchmarking approach, based on a published strategy, to compare the performance of several popular shape-based approaches with each other, and with a non-shape comparator. This allows us to determine the ability of the developed descriptors to identify active compounds. Accordingly, it allows for an analysis of whether or not the developed shape method correlates with on-target activity.

3. We carried out two prospective studies, assessing the global and local shape descriptors developed, to determine whether either could suffice to identify novel compounds that would not have been chosen with traditional approaches.

4. We conducted a detailed analysis of one of these prospective studies, determining in-depth the relationship between ligand and pocket shapes for a particular target.

5. We attempted to construct a model to translate from a target pocket back to a complementary ligand, to assess whether the geometry of a protein pocket contains sufficient information to specify that of a binding ligand.

Chapter 3

Methods

3.1 FRACTVS Package

The FRACTal Virtual Screening (FRACTVS) package allows for the calculation of the correlation dimension (D_C) of FD (henceforth referred to as FD) for small- and macro-molecules. Implementations for both global and atomistic FD are provided, and utilised for retrospective and prospective studies.

Here, we aim to highlight and discuss snippets of the code to provide a detailed overview of the package developed. Chosen functions cover the basic operation of the software, algorithmic considerations and optimisations which aid in reducing the overall computational complexity of the approach, and provide a broad overview of the command logic used, without entering into exhaustive detail with regards house-keeping tasks, simple transformations, and trivial optimisations, etc. A complete version of the software, along with documentation, will be provided in the public domain upon publication, and is available in the first instance from the Schneider lab. To facilitate this discussion, clearly-marked pseudocode may be used in place of the working code. The code-base is written in Python 3.7 [361], with a Python 2.7-compatible version available. Python is an object-oriented scripting language, with considerable community support, and a wide variety of packages available to facilitate easy prototyping.

Many open-source packages have been used to aid the process of development. RDKit provides much of the chemical logic for the software, facilitating washing procedures, keeping track of molecular properties and geometry, and providing for the minimisation of small molecules.

Numpy consists of a library of highly-optimised numerical subroutines, with a Python interface, allowing speedy and memory efficient distance calculations. SciPy is used in a similar fashion, likewise offering efficient pairwise-distance calculations. NanoShaper facilitates rapid and accurate calculation of molecular surfaces for small- and macro-molecular objects. BioPython provides a suite of useful tools for the processing of PDB files into formats more amenable for inclusion in the developed pipeline. Multiprocessing provides an easy interface for distributing jobs on multi-core machines. Finally, the Cython interpreter, and numba package, were used during development, to facilitate the inclusion of fast C subroutines.

Our approach is concerned with the description of a set of vertices placed on the Solvent-Excluded Surface (SES) surface of a given molecule. The SES, or Connolly surface, was chosen for a mixture of chemical and algorithmic reasons; briefly, that it presents a more 'physical' surface than the Solvent-Accessible Surface (SAS), that it has been observed to be more useful for geometric interactions, and that the prominent invaginations seen with the SAS are problematic for our algorithm, given that they distort the local density of the surface. The rationale for this choice of representation is expanded on in subsection 3.1.2. To calculate the Connolly surface, we employ the software NanoShaper [362], which has been shown to be orders-of-magnitude faster than MSMS [363], at the cost of a somewhat-increased memory footprint. In addition to its speed, scalability, and improved reconstruction on some tasks, it also exposes some functionality for the segmentation of its vertex set according to their corresponding atom; this allows for a simple definition of a surface exposed atom as one with at least one associated vertex, and for the consideration of atom-specific local FD environments (subsection 3.1.5). We use the standard radius value for an idealised model of a water molecule, 1.4Å [115].

As discussed by Grassberger and Procaccia [200], the correlation exponent provides a good approximation of the analytical solution for fractal dimensionality in non-pathological cases. We employ a version of this algorithm optimised for our particular use-case in small- and macro-molecules, as alluded to previously. The correlation exponent (and thus FD) is obtained by observing the variation in an unbiased estimator, the correlation sum $\hat{C}(\delta)$, with δ, a distance in euclidean space, in D, where $D = < \delta_{min}, \delta_{min} + \delta_{int} \ldots, \delta_{max} >$. For a given configuration, δ_{min} represents the minimal value of δ considered, δ_{int} the step-size, and δ_{max} the maximal value. The algorithm is described in depth in section 1.2.7.

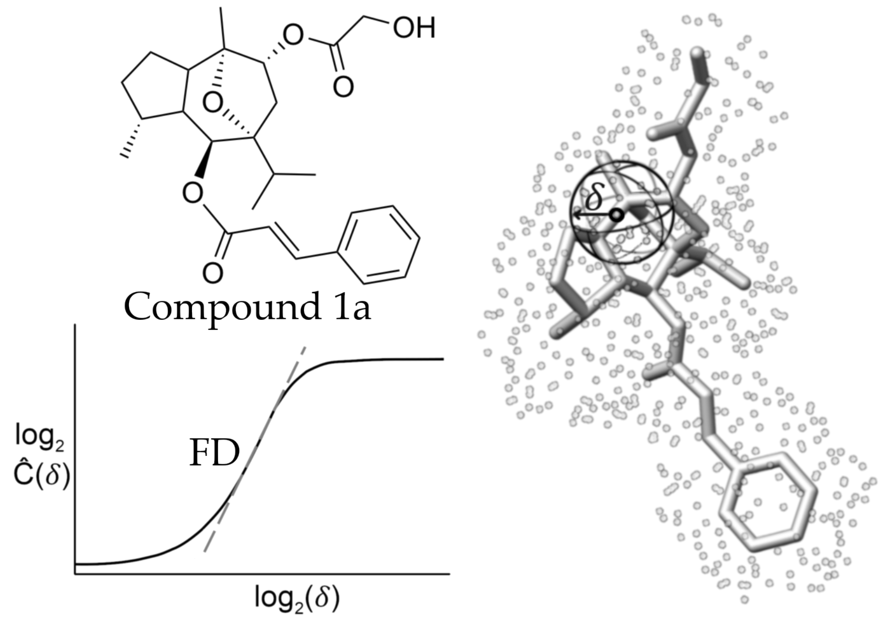


Figure 3.1: Schematic overview of the FD calculation procedure, as defined by Grassberger and Procaccia [200], and applied to the natural product (-)-Englerin A. Given a simple 2D representation, we generate a range of conformers, describe them in terms of a set of vertices on their Connolly surface, and characterise the relationship between these points in terms of the gradient of the log-log plot of how many pairs are within a distance, δ, as we vary δ. This gradient gives us an estimate of FD, more specifically the correlation dimension, D_C.

We construct the distance matrix to be evaluated on an atomic basis, to reduce the computational complexity of the problem, and to enable FD measures based on the surface local to each atom in a molecule, which we term atomistic FD (AFD). To enable this, we extract the atom centroid co-ordinates and identities from the RDKit molecular representation, construct a Euclidean (L2-norm) matrix of the inter-atomic distances using functionality exposed in SciPy (v 0.18.1), and subtract δ_{max}, the sum of the atomic-radii vector, A, its transpose, AT, and a small corrective factor (2.8Å, predicated on the diameter of the probe molecule employed in NanoShaper) (see Figure 3.3). Pairs of atoms whose value in this matrix are ≤ 0, and where i$\neq$j, are then deemed to be 'within-range' of one another. For each of these pairs, their corresponding set of vertex-coordinates are retrieved from the dictionary discussed in subsection 3.1.2, and a Euclidean matrix constructed.

In addition, an intra-atomic matrix is constructed, which considers the relations between points attached to the parent atom. These two matrices are then considered under a histogram rule, which bins the distances with the edges defined in D, and the corresponding counts are returned to a global per-bin counter. This allows for one-step evaluation of the Heaviside step function for all δ in D , and also a reduction of complexity by narrowing the problem space, which is of considerable value in the case of macromolecular analysis. The equation of the line is then obtained by means of a Theil-Sen[364] estimator, which is reasonably robust to outliers. A schematic overview of this process, moving from molecule, to surface, to dimensionality estimation, is given in Figure 3.1. Utilising this atomic-level FD information, we then proceed to build fingerprint representations, using the ECFP approach [43] previously described in the literature, to construct a Fractal Dimensionality FingerPrint (FDFP) for small molecules, and a modified approach to facilitate a similar representation of macromolecules, as per the E3FP-NoStereo algorithm described in the literature [106]. More detail on all aspects is given in the following sections.

Comparing our approach to previous work, we see some general, and specific, differences. The early work of Lee et al. [115] established the utility of FD for considering binding propensity. Their method, utilising area over probe size, gives a similar range for average FD to those seen in later work [169], but their method (spherical projection) for summarising the regional variation in the surface results in artefacts, as discussed by Pettit et al. [164]. While the latter authors use the same basic FD definition as Lee et al., they consider it as a local property, building a smoothed AFD representation. This averages the roughness over a patch of the surface by considering the relationship between area and probe size for all of the surface within 5Å of each atom. This approach, considering local roughness as an important property, was extended by Todoroff et al. [169], who emphasised the importances of a pattern of smooth and rough local patches in defining a binding site. Again, they consider this as an atomistic property. Our initial approach, calculating FD on an atomic basis, has several features in common with the work of Todoroff et al. [169]. However, their implementation, refinement, and the direction taken with validation, results in a substantially different representation. On an algorithmic basis, several innovations and efficiency improvements have been made, substantially reducing the computational overhead, and allowing for a more rigorous estimation of FD (subsection 3.1.5), which is less sensitive to parameter choice, and therefore more easily extensible to new problem domains. Importantly, we then re-integrate these local-similarities into a fingerprint structure, which allows for the easy comparison of small and macromolecules on a pairwise basis, and facilitates the prospective studies undertaken. This allows for a representation which combines both the very localised, atom-level roughness description, and enables consideration of the shape of binding pockets at patch or pocket level through combination of AFD with a geometric fingerprinting approach. Equally, it allows for the useful application of the FD approach to capture information about small molecules, and to model the relationships between small- and macro-molecules in detail, based on their geometric complementarity.

3.1.1 calculate_correlation_exponents.py

This interface module offers various means of calculating global and fingerprint FD for small and macro-molecules, supporting command line usage, and import for use in another script. In general, this module orchestrates the operation of the other major modules, containing mainly command and output code, and is the most likely start-point for an end-user. Given an SDF file containing small molecules, this script will note the desired maximum number of conformations, δ_{min}, δ_{int}, and δ_{max} values. All can be passed as user-defined parameters, but sensible parameters based on empirical studies are set by default. Text files with SMILES, and Mol2 files, are also valid inputs. For PDB files, the result is slightly different, as the poses for a ligand in the PDB file are deemed canonical. Washing of the protein and any associated ligands is still enabled, as there are often issues with PDB files, whether publicly-deposited or privately-generated [365], and we wish to obtain a consistent representation for our further calculations. Initially, for both small- and macro-molecules, the approach is similar: The number of unique entities is determined, and chunked across the available processors. This is to avoid the computational overhead of 'spinning-up' a new process for each molecule, which becomes excessive for large numbers of small-molecules. This requires the input file(s) to be processed using the functionality exposed in the *MolecularRepresentation* class, which will be discussed below. Following this set-up process, which, broadly speaking, involves checking that molecules are comprehensible and valid, individual processing takes place. The *MolecularRepresentation* instance mentioned earlier is utilised again, to generate conformers if necessary, and a molecular representation for all molecules associated with a given parent. Following this, the script iterates over each molecule in the parent object, determines whether the last step was successful, and, if so, calculates global and atomistic FD values for each. Leaving the code-block, eventually the results are collated, associated with their parents, and written as SDF files, or as log and matrix files for PDB input. These can then be further processed with the *fingerprinter.py* script, producing fingerprints for small molecules, and any associated near-molecule protein pockets in the case of those extracted from PDB files.

3.1.2 molecular_representations.py

The processing of raw input files into a format suitable for FD calculation is handled within this module. Given an input file, this script attempts to determine the input file-type from its MIME type and suffix, and then either refers it to *PDBProcess.py* for processing, if detected to be a PDB file, or constructs RDKit mol objects from the list of input SMILES, Mol2 or SDF file provided. These are then washed (subsection 3.1.3), and the conformer generation script described in subsection 3.1.4 and as outlined in listing 3.1 is called, if enabled.

Having washed our input molecule and generated a suitable conformer, or obtained such geometric data from a crystallographic source, the issue of molecular representation remains. Two commonly adopted surface definitions, from the fields of structural biology and molecular visualisation, are the SAS and SES. The former of these, also known as the Lee-Richards [115] or Shrake-Rupley [116] surface, describes a molecule in terms of a surface defining positions where a solvent probe, commonly water, may be positioned most closely to the molecule. This can be thought of, and implemented as, building a surface based on rolling the probe over the molecule, and tracking the coordinates of the centre of the probe throughout, as can be seen in Figure 3.2. This is a useful representation, computationally tractable, and with a solid basis in structural biology, having been developed originally to facilitate protein-folding efforts [366, 367], and analysis of hydrophobic effects in free-energy transfer [368]. Essentially, it describes the potential cavity that a given molecule would form in bulk solvent. As can be seen, it functions to a certain extent as an ‘expanded’ van der Waals surface, where the expansion is equal to the probe radius, preserving the invaginated surfaces between atoms.

An alternative approach, developed by Connolly et al. [52, 117, 118], is to consider another property of the probe molecule interaction. If we instead take the contact surface between the probe and molecule, we

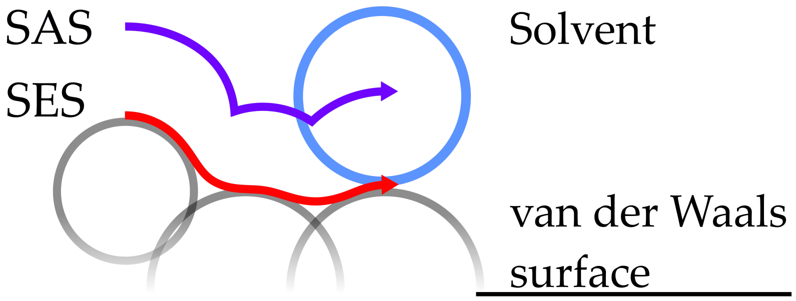


Figure 3.2: Schematic representation of the solvent-accessible [115, 116] (SAS, purple), solvent-excluded [117, 118] (SES, red) and van der Waals (grey) surfaces of a molecule. A representative solvent molecule is included in blue. When 'rolling' this solvent probe over the van der Waals representation of a molecule, the solvent-accessible surface can be defined by considering the area delineated by the centre of said probe molecule. The solvent-excluded surface can be defined as the collection of points of nearest contact, i.e., the point of closest approach of the solvent probe to the molecule, for all points on its van der Waals surface. As such, it consists of contact areas, where it is identical to the van der Waals surface, and re-entrant surfaces, representing volumes not occupied in the hard-sphere molecular representation, but which are too small to permit the entry of the solvent probe.

elucidate the so-called SES, as depicted in Figure 3.2. This approach results in a smooth envelope, 'wrapping' the molecule. Although initially intended primarily as an aid to visualisation of interaction surfaces, this representations gained widespread use in the field, representing a slightly more intuitive depiction of molecular volume than the SAS. As discussed in the literature [369–371], the latter is a somewhat 'unphysical' representation for many purposes, preserving the invaginations projected outwards from the van der Waals surface, and has proven less suitable than the SES for surface-matching and geometric docking of binding partners [193, 372–375]. Given that the SES has been identified as a better option for visualising and predicting interactions on a geometric basis, and that it is somewhat more physical, we adopt it here. Additional considerations, from the perspective of the algorithm here employed, relate to the acute invaginations seen with SAS, and with analysis of binding partners. With the former, these represent areas of substantial density, in terms of number of vertices per Å^3, owing to the inwards folding of the surface, and as such are over-weighted compared to solvent-exposed surfaces by our FD calculations. The latter point overlaps with observations made about the relative utility of these representations for geometric docking, in that surface-clash is to be expected given the addition of the probe-radius to the van der Waals surface, especially given the certain degree of noise often present in crystallographic structures, rendering FD calculation of the binding pocket for a given molecule problematic.

To generate the Connolly surface for a given small- or macro-molecule, we use an external software, NanoShaper [362]. This uses a ray-casting approach to rapidly, and accurately, locate the SES. An exhaustive list of parameters chosen is included in two parameter files included in the repository, namely *ligand_nanoshaper_config.prm* and *protein_nanoshaper_config.prm*. Briefly, there is substantial overlap between the parameter files, barring some adjustments made to allow for a higher memory footprint for macro-molecules. Otherwise, we specify 'xyzr' input, ask for the resulting vertices set to be divided on a per-atom basis, set the radius of the solvent probe to 1.4Å, the vertex density to to 7Å^{-2} and the proportion of the ray-casting box occupied by a given molecule to 50%. Additionally, we enable 'accurate triangulation' of surface vertices, a post-processing step which seeks to verify the placement of vertices on the traced surface, and which was noted by the authors to significantly improve stability with minimal computational overhead. This module writes the necessary input and configuration files for NanoShaper, triggers its operation, and, depending on the outcome, processes the resulting files into a format amenable for further calculations. The resulting data structure consists of a dictionary whose keys are atom indices, and values are numpy arrays of vertex co-ordinates. In addition, a second dictionary mapping atom index to atom position (as defined in the coordinate list produced by minimisation, or loaded from PDB file) is generated.

In the case of a PDB input file, this script also selects all residues with a surface-exposed atom, i, whose minimal distance, d, to any atom, j in any associated ligand, is $\leq$ 3.5Å$+(r_i + r_j)$, where r_i and r_j represent the van der Waals' radii of atoms i and j respectively. The threshold value was chosen by consultation with literature on typical interatomic bond distances from crystallographic databases [376], coupled with empirical observations on the resulting bit density of generated protein pocket fingerprints. We observed that choosing lower thresholds, such as 2.5Å resulted in too-few contact atoms in many instances, whereas increasing to our current threshold avoided adding many fewer indexed atoms in cases where the lower threshold had proven adequate.

3.1.3 Washing procedure

Using an approach adapted from several sources (RDKit cookbook, MolVS project, Sanifix project), we perform a basic washing workflow for input small molecules. This procedure is optional, but the results of conformer generation, and subsequent FD description, depend to a large extent on having a consistent basis. The tasks performed by this script can be divided into fragmentation, neutralisation, and correction. The former category contains two functions, one of which disconnects covalent organic-metal bonds under certain conditions (defined by a pool of curated SMARTS), the other attempts to remove salts, as these do not contribute to the minimisation step. Neutralisation consists of SMARTS patterns for the modification of imidazoles, amines, carboxylic acids, thiols, sulphonamides, enamines, tetrazoles, sulphoxides and amides such that their charge state is consistent. Correction contains a SMARTS pattern to correct common issues with aromatic nitrogens, and charge correction, which checks whether each atom in the molecule has an appropriate charge, adds implicit hydrogens (necessary for minimisation), and performs molecular sanitisation.

This final step proceeds as follows: first, valence states are checked, and a few common, non-standard valence states are corrected for. These include forcing some charged groups into a consistent zwitterionic form. Then, implicit and explicit valences are re-checked, the molecules are kekulised, and radical electrons assigned. Aromaticity flags on each atom are re-checked, as are bond conjugation and chirality, before finally re-checking whether all implicit hydrogens have been made explicit. If RDKit-based conformer generation is enabled by the user, the washing procedure will also remove those molecules containing atoms for which there are no available MMFF parameter sets. To the best of our knowledge, no open-source software implemented in Python is available for tautomer generation and pH-appropriate charge assignation, although the MolVS package is a considerable contribution in that direction.

For macromolecules, we utilise the PDBFixer [377] for standardisation purposes, along with a custom script for chain nomenclature standardisation. This consists of residue nomenclature standardisation, selection of a single position for atoms with multiple positions listed in the crystallographic file, removal of hydrogen atoms, removal of crystallographic additives (glycerol, magnesium, salts, etc. - for a complete list, please see *PDBProcess.py*).

3.1.4 Conformer generation

The generation of biologically-relevant conformers for small molecules remains problematic. The reproduction of crystallographically-obtained ligand conformations remains the gold standard in assessing the performance of conformer generation approaches, but is problematic [378]) given the dynamics of crystal formation (with especial reference to induced changes in side-chain conformations, leading to modified ligand poses, with different space groups) [379], and artefacts introduced when such crystal poses are obtained by soaking [380, 381]. A common solution, adopted for a variety of shape-based approaches published to date, is to simply incorporate multiple conformers for each molecule when performing a similarity comparison. Our strategy for generating a small set of diverse, representative, and chemically reasonable conformers is based on attempts to optimise this process for use with the ROCS shape-based virtual screening software [48, 263, 268, 382], whilst utilising open-source chemoinformatic and force-field implementations and the best practices associated with each [383]. To maintain easy compatibility, and to facilitate deployment, we utilise the conformer generation capabilities presented in the RDKit, which has the advantages of a well-described implementation, and comparable speed to commercial competitors [383]. A brief overview of the code to achieve this is provided in listing 3.1.

Snippet 3.1: molecular_representations.py - conformer generation

```
PSEUDOCODE

def generate_conformers(mol):
return rdkit_geom_optimisation(mol)

def embed_molecule(self, mol, n_conformers=10):
params = AllChem.ETKDG()
params.pruneRmsThresh = 0.5
params.useBasicKnowledge = True
params.useExpTorsionAnglePrefs = True
params.useRandomCoords = True
params.numConfs = max(10, n_conformers)

cids = list(AllChem.EmbedMultipleConfs(mol, confnum, params))
conformers = list(mol.GetConformers())

for conf in cids:
valid = check_if_valid()
if not valid:
remove_conf(conf)

return success, mol

def rdkit_geom_optimisation(self, mol, mode, n_conformers):
success, mol = self.embed_molecule(mol, n_conformers)
if not success:
return None, None

optim_status = AllChem.MMFFOptimizeMoleculeConfs(
mol,mmffVariant='mmff94',maxIters=1000
)

min_energy = min(energies)
conformers = list(mol.GetConformers())]
conformer_properties = calculate_energy_delta(conformers)

for idx, [conf] in energy_sort(conformers):
if energy <= min_energy + 5:
keep(conf)

return mol, conformer_properties
```

For initial conformer generation, we use the Experimental-Torsion basic Knowledge Distance Geometry (ETKDG) [384] implementation provided in RDKit. This utilises experimentally-observed information on torsion-angle preferences from a database of crystallised small molecules to facilitate rapid generation of reasonable conformers, using an augmented distance-geometry approach. In addition, incorporating explicit additional knowledge terms for plausible ring geometries was observed to facilitate reproduction of crystal conformations with fewer poses than purely distance-based minimisation approaches. In accordance with our literature survey, we generate 10 conformers at this initial stage, with an RMSD $\geq$ 0.5Å to all others in the set, setting random coordinates on initialisation for each iteration. This threshold value matches the significance threshold employed by Riniker et al. [384] to determine whether the bioactive conformation was replicated within reason, and, as such, is used here to meaningfully distinguish between conformers.

A second threshold value obtained from that survey, and a means of constraining the number of generated conformers while maintaining adequate coverage of the conformational space, is to impose an energy window of 5 kcal/mol [385], meaning that all conformers with an energy greater than the sum of the lowest energy plus this threshold value are discarded. This entails calculating the energies of the conformers under a forcefield. The Merck molecular forcefield 94 [**halgren˙1996**, 386–388] (MMFF94) as implemented in the RDKit toolkit [389], and as utilised in a modified form (without the terms capturing electrostatic and van der Waals interactions) in the popular OMEGA conformer generation software [390], has been employed in previous evaluations and applications of shape-based virtual screening approaches [87, 100, 385]. It was originally parametrised using the composite quantum chemical method HF/6-31G* [391], and has shown considerable utility in the generation of pharmacologically-relevant conformations, although Hawkins et al.[390] report that it overestimates strain contributions, leading to failure to reproduce crystallographic conformations, primarily in cases where a molecule has many degrees of freedom.

Following this we then remove explicit hydrogens added in subsection 3.1.3. Atomistic FD calculation is only carried out for heavy atoms, owing to the high degree of rotational flexibility of hydrogen observed during minimisation.

3.1.5 fractal_calculators.py

This module performs all calculations required to estimate the fractal dimension of a given molecule, at both the global and atomic levels. As discussed in section 3.1, a trivial implementation of the algorithm for calculating the correlation exponent, and approximating the fractal dimension, of a point cloud would be to calculate the pairwise distances between all vertices, V, and, from there, to analyse V for δ in D.

In the course of developing the AFD description, the basis of the FDFP method, we modified this initial approach to take advantage of the underlying structure present in our case, to wit, the chemistry defining relations between atoms, and, therefore, shaping the molecular surface. Our approach is conceptually similar to that of Theiler [392], barring that in our case, we utilise the geometric information implicit in our conformers rather than choosing an arbitrary grid structure. We utilise the dictionaries described in **subsection 3.1.2, molecular_representations.py** to define which sets of vertices should be compared, on the basis that atoms which are too far distant will not have any vertices within d_{max} of one another (see: Figure 3.3), and therefore will not affect the final FD value.

$$C = \theta(M - R - R^T + R_S)$$

where C is a logical matrix encoding the homogenous relation over the set of atom indices defined by θ, M is the matrix of interatomic Euclidean distances, R the van der Waals radii of the molecule's atoms, and R_S a solvent adjustment factor, fixed to 2.8Å, based on the diameter of the solvent probe utilised for NanoShaper. θ is the Heaviside step function,

$$\theta(x) = \begin{cases} 0, & x > \delta_{max} \\ 1, & x \leq \delta_{max} \end{cases}$$

For all pairs of atom indices, $(a, b) \in C$, we compute the squared Euclidean distances between the vertices belonging to each atom, and bin the result using a cumulative histogram function. Here, the bin edges are defined as per the vector of delta values D previously described, taken to their square, to avoid the computationally-expensive square root operation. The global correlation sum is obtained using the atomistic contributions, albeit restricted to the sum of histograms for the multiset of unique unordered pairs in C. The histogram process is identical to applying the procedure described in section 1.2.7, for each δ in D.

Snippet 3.2: Fractal_calculators.py

```
PSEUDOCODE
def calculate_fd(atom_coordinates, atom_radii,
vertex_coordinates, dvector):

interatomic_dist_mat = euclidean_distance(atom_coordinates)
interatomic_dist_mat -= atom_radii
interatomic_dist_mat -= atom_radii.Tminipage
dmax = max(dvector)

valid_pairs = list(zip(np.where(interatomic_dist_mat <= dmax)))
pairwise_distances_dict = {}

for i_idx, j_idx in valid_pairs:
        if i_idx > j_idx:
                pairwise_distances_dict[i_idx][j_idx] =
                pairwise_distances_dict[j_idx][i_idx]
        else:
                pairwise_distances_dict[i_idx][j_idx] =
                euclidean(vertex_coordinates[i_idx],
                vertex_coordinates[j_idx])

global_fd, atomistic_fd = [], {}

seenx_pairs = []
for i_idx, pairwise_dists in pairwise_distances.items():
        for j_idx, ds in pairwise_dists.items():
                binned_vals = histogram(ds, bins=dvector)
                if (i_idx, j_idx) not in seen_pairs:
                        global_fd += binned_vals
                        seen_pairs.append((i_idx, j_idx))
                atomistic_fd[i_idx] += binned_vals

        atomistic_fd[i_idx] = theilsen(
                log(dvector),
                log(cumulative_sum(atomistic_fd[i_idx])))

global_fd = theilsen(log(dvector),
log(cumulative_sum(atomistic_fd[i_idx])))

return atomistic_fd, global_fd
```

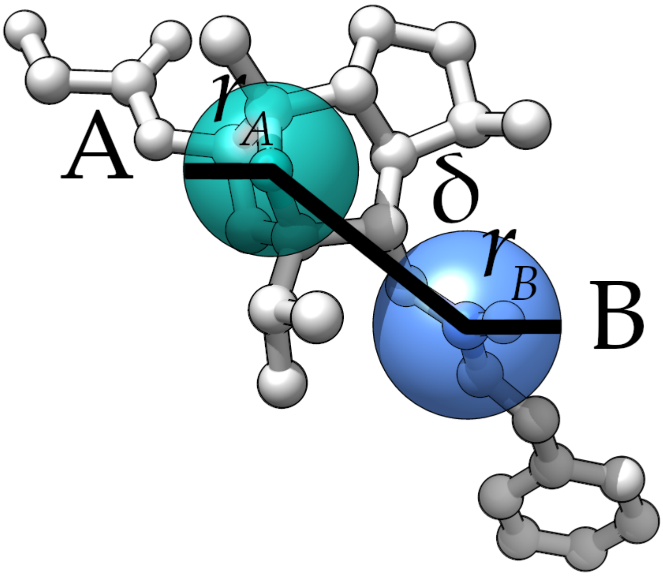


Figure 3.3: Illustration of molecule-specific adaptations to the correlation integral algorithm, to improve run-time and memory performance whilst retaining standard accuracy. We calculate the Euclidean interatomic distance matrix M for all pairs of atoms in a given molecule, based on the coordinates of the centres of their representative atomic volumes. For atoms A and B, in green and blue respectively, the distance between these centres is δ. The van der Waals radii of the two atoms, r_A and r_B , are subtracted from this distance to a give the adjusted interatomic distance, the closest that two points placed on the representative spheres of the atoms can be, given their radii and the distance between them. Finally, this is adjusted with a parameter for the solvent radius, adding 2.8Å in each case, and then transformed with a Heaviside step function, evaluating to unity for only those atom pairs whose adjusted distance is less-than or equal-to the specified δ_{max} threshold, giving a binary relation matrix, C, describing those pairs of atoms which are likely to have patches of their surfaces within distance δ_{max} of one another. As a result, the speed of the FD calculations depends on δ_{max} to a large degree. This adjustment to the basic algorithm is described more formally in section 3.1.5, fractal_calculators.py.

The AFD approach differs slightly, in that for a given atom, A, we would consider only those elements of V corresponding to the indices of the set of vertices for A. Thus, it is a reflection of the distribution of the surface from the perspective of the surface of each template atom, rather than of the pairwise distances for the surfaces 'belonging' to each nearby atom. This is an important distinction, as this allows us to retain arbitrary addressability of individual portions of the surface, whilst allowing for the consideration of local behaviour.

Having determined the correlation sum for both the global and atomistic cases, obtaining the exponent is accomplished through use of a regressor. As discussed in section 1.2.7, the use of fractal conceptions of dimensionality to describe non-fractal objects encounters some difficulties, one of which is that non strictly-self-regular objects result in irregular correlation exponents, as the rate of repetition is scale-variant. To reduce the impact of this, we chose to utilise the Theil-Sen estimator, a robust regression method with high noise tolerance which returns identical results to least-squares regression in normal cases, but can handle the addition of noise (up to 29.3% of total data points [393]) without producing markedly different estimates of the exponent.

3.1.6 fingerprinter.py and graph_construction.py

As discussed in section 1.2.4, performing pairwise shape comparison of large sets of molecules in a reasonable time-frame is difficult. Several approaches to avoiding costly pairwise alignment of molecules have been implemented, often featuring a vector representation of molecular shape. To compare the performance of our global FD (GFD) and AFD approaches, we sought a means of encoding the atomistic information in a fashion which would permit rapid comparison whilst retaining a physical grounding. We chose a fingerprint approach, based on the flexibility and ease-of-use associated with this molecular representation. The popular ECFP method [43, 394] allows for the incorporation of per-atom labelling information. In the original definition of ECFP, the daylight atomic invariants [395] are used to label each atom, namely; number of heavy neighbours, atomic number, atomic mass, atomic charge, and number of attached hydrogens, extended with a label signifying whether a given atom is a member of at least one ring.

Without reiterating the algorithm in detail, these labels are then hashed, returning an integer value, and relabelled with the transformed label. In subsequent iterations, the labels of topologically-connected atoms are considered, concatenated to the current label, and the result is then hashed again. This process is continued for a fixed number of iterations, which provides the suffix for the ECFP prefix, e.g. ECFP4 indicates that two such cycles are completed, where two is the radius, doubled to give the diameter of four referred to. All of the hashed labels generated during this process are then used to generate a bit vector representation, where the on-bits are specified by the modulus of the label over the desired length of the resultant bit vector.

We re-implemented the algorithm described in the original paper, swapping the set of invariants employed there for a discretised version of the AFD, where we round the calculated value to two decimal points, and then multiply the result by a factor of 100. This discretisation step was introduced to avoid the problem of small differences in float values having an outsized impact on the resulting bit allocation, with two decimal points being chosen such that all resultant numbers would fit naturally into the commonly used 1024-bit vector, and based on observed variations in AFD between runs with identical conformers, due to some slight variation in the surface generated by NanoShaper. Using this discretised representation to replace the invariants specified in the earlier implementation of the ECFP algorithm gives us the final FDFP vector.

Generating topological fingerprint representations of macromolecules requires a different approach. Many of a surface-exposed atom's neighbours are buried, and therefore have no associated AFD value, meaning that their atomic invariants vector would be null. In addition, two atoms in close proximity from a geometric perspective, may have a very large topological distance, given the intricately folded nature of proteins, for example. To circumvent this, we use the subset of near-ligand, surface-exposed atoms whose identification is described in subsection 3.1.2, molecular_representations.py, and generate a fingerprint adapted from the E3FP NoStereo variant [106] described by Axen et al. Briefly, this approach differs from the topological fingerprint in that it uses concentric rings in $\mathbb{E}^3$, centred on a given atom, to define neighbourhoods, rather than topological distances obtained from walking along the molecular graph. In their original implementation, the authors use the same set of invariants as described for ECFP, above, and include information on bonds, unbonded atoms, and stereochemistry. The algorithm proceeds in the same fashion as ECFP, albeit increasing the geometric discrete distance, rather than the topological, with subsequent iterations. These distances are drawn from a set of shells where the radius increases by 1.718Å upon each iteration. For our implementation, we ignore specific bonding information and stereochemistry, and treat all atoms in the surface-exposed set as unconnected. Again, we replace atom invariants with the discretised AFD described above. The authors of that study note that the addition of stereochemical information did not improve performance of E3FP in their chosen test, likely indicating that such information is captured implicitly.

3.1.7 PDBProcess.py

This module is primarily concerned with the processing of PDB-format files to facilitate FD calculation for small molecules. It removes common crystallographic additives, such as polyethylene glycol, imidazole, and calcium ions, removes chains for which no bound ligand is in range, and extracts ligands. Additional functionality for providing a user-specified binding site is also provided, but disabled in normal usage. Specific functionality for the processing of scPDB data is included: mainly this involves incorporating ligand files, provided in the Mol2 format, into the protein PDB file, to enable smooth integration with the established PDB workflow, but also allows for an alternative binding site mode, in which the definition is obtained from the scPDB database rather than using the rules described in section 3.1.2, molecular_representations.py. In addition, this module provides some convenience functions for retrieving the corresponding UniProt IDs for a given PDB ID, along with any associated affinity data stored in the BindingDB database.

3.2 Benchmarking study

We adopted and adapted the benchmarking study by Riniker and Landrum [214]. Whilst the code provided (bit.ly/rdkitbenchmark) was very useful, and easily extensible with additional 2D descriptors, substantial modifications were required to facilitate inclusion of shape-based approaches, for reasons of computational efficiency, inclusion of logic for multi-conformer comparisons, and the extraction of additional endpoints for extended analysis of the retrospective results.

3.2.1 Datasets

The merged dataset consists of actives and 'decoys' (presumed inactives) from the benchmarking study platform for 88 targets; 17 from MUV [235], 21 from DUD [232], and 50 from the ChEMBL subset. For a description of the target set properties, please see the original paper [214].

The MUV approach is based on data extracted from the PubChem bioassay database. Data from pairs of analyses (one high throughput for hit-finding, one low-throughput for hit-verification) were sourced, extracting active and inactive compounds for a set of 17 targets. Compounds which aggregated under test conditions, promiscuous binders, and optically-active compounds were filtered out. As noted by the authors, their approach is not sensitive to false negative data, as compounds which register as negative on the HTS screen were not progressed for the high-quality screen. The main innovation in their approach is to consider the embedding of active data in the validated decoys, by removing active compounds from further consideration which are very far from, and have an insufficient number of nearest-neighbours in, decoy data sets under the molecular representation chosen (a vector of simple molecule properties, such as atom counts, acceptor and donor counts, LogP, chiral centre counts, and number of ring systems). Two approaches are taken, one to separate active compounds in the descriptor space by an equal amount, and another to surround these with decoys

One issue with this dataset, as discussed previously in the literature [396], is its treatment of so-called 'activity cliffs', where very similar compounds can have markedly different activity values. As such, it is hard to know whether a method has retrieved a molecule in the right chemical space, where a slight reconfiguration could markedly boost activity without much superficial alteration to the molecule. By assessing compound pairs which had been trialled on other targets, they noted that this cliff-like behaviour correlates with the variation in per-target performance noted in their benchmarking study.

DUD is a very popular resource for the comparison of different docking methodologies. It is constructed so as to generate a set of decoys for a pool of known actives with very similar physical properties, whilst retaining distinct chemistry. It is based on a curated set of crystal structures, associated ligands, and a set of decoys chosen from the ZINC database on the basis of sharing some physical properties (Molecular weight, Hydrogen-bond acceptors and donors, LogP and number of rotatable bonds), whilst having a maximum equivalent Tc of 0.7 under the Daylight fingerprints (equivalent: the authors used CACTVS fingerprints, with a threshold of 0.9). In the years following its release, the set was very actively adopted, leading to some focused criticism of the methodology underlying its construction, such as poor treatment of charge, and that some of the annotated decoy compounds have later been shown to have on-target activity, leading to further enhancements [239] and the development of related resources [397].

The original authors note that it is, by definition of its construction, fundamentally poorly-suited as a means of discriminating between ligand-based approaches, whether 2- or 3D [236], as the active sets for each target contain trivial analogues, readily-separable by even the simplest methods. Regardless, it remains a commonly-used benchmark for LBVS. As such, we have retained it here for comparison purposes. The subset utilised in the benchmarking study consists of 20 targets with 30 or more annotated actives. We excluded FXa from further consideration, as a high proportion of its annotated active compounds were invalid, as previously reported in the literature [397].

The ChEMBL[243, 398] database reflects the efforts of decades of research and innovation in medicinal chemistry. As such, it contains records on $\geq$1.6 million compounds, and over 9,000 target proteins, approximately 4,000 of which are human. It is constructed primarily by manual data curation, alongside data submission by companies and collaborative research efforts [399]. Given the multiplicity of data sources, annotation quality varies from set to set; this is accounted for with confidence measures. Earlier efforts to construct benchmarking sets from ChEMBL [243] utilised this to facilitate automating the process somewhat, choosing targets for which at least 50 active compounds were annotated, at the highest confidence level, and with $\leq$10 µM potency. This was repeated for the benchmarking approach detailed here. Additional filters, such as removing molecules with a molecular weight $\leq$700 gMol^{-1} or which contain metal ions, were imposed. Finally, a diversity picker was used to pick the 100 most-diverse actives for each target. Decoys for this set were constructed by randomly selecting two compounds from ZINC with an ECFC0 Dice similarity $\geq$0.5 for each active.

3.2.2 Benchmarking study workflow

The implementation of the workflow described by Riniker and Landrum is included in **gfd_fdfp_analysis.py**. This script handles all implemented features from that study, although generation of conformers and FDFPs are managed separately. To assess the impact of the δ_{min}, δ_{max} and δ_{int} parameters, and the role of the fingerprint radius, a script which generates the necessary files for all combinations is utilised. This also handles the necessary alterations for each run, to ensure that the parameters passed to the benchmarking script itself are appropriate.

3.2.3 Ligand-based screening methods

Extending the work of Riniker and Landrum, we incorporate their initial findings to narrow the pool of 2D descriptors, whilst including several 3D approaches, including that described in this manuscript, and several established methods. ECFP-type descriptors are included to provide a baseline performance level, firstly, as they were found to provide the best overall performance for 2D methods in the original benchmarking study, and, secondly, owing to their continued popularity in virtual screening and as inputs for machine-learning studies. The USR, and USRCAT, 3D descriptors have a very light computational footprint, and offer a comparison to other alignment-free shape-based screening methods. OpenEye Scientific's ROCS 3D tool remains the gold-standard approach for shape-based virtual screening, and has been adopted widely in academia and industry for the identification of novel ligands. USRCAT and ROCS colour benefit from the addition of explicit pharmacophoric and electrostatic information, which was observed, in each case, to substantially improve performance in ligand retrieval over their baseline variants. A full overview of these approaches is given in section 1.2.4. N.B. All future mentions of ECFP4 and ECFP6 refer to the Morgan fingerprint implementation provided in the RDKit, unless otherwise specified.

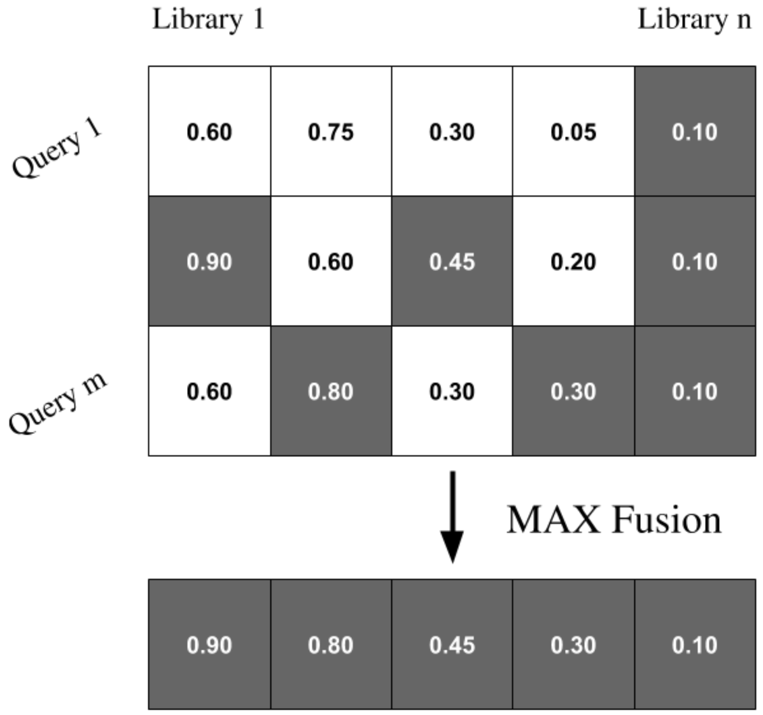
Table 3.1: Overall number of compounds per database, and resultant conformers generated after a process of de-duplication.

Database	**Initial count**		**Unique count**		**Conformer count**	
	Actives	Decoys	Actives	Decoys	Actives	Decoys
ChEMBL	7834	10000	6739	10000	32572	53777
DUD	1917	89309	1617	61037	8056	330739
MUV	510	254999	492	96340	2492	511914

Data generation and processing For the sake of comparability, multiple conformers are generated once for each set of actives and decoys described. The number of actives and decoys per database, along with the number of unique molecules and conformations generated are included in Table 3.1. On average, approximately five conformers remained for each unique molecule after conformer generation in the manner described in subsection 3.1.4, Conformer generation. For each combination of parameters, we carried out FD calculation and fingerprint generation as described in section 3.1. As the other methods trialled do not depend on these parameters, benchmarking is only performed once per dataset for the other virtual screening methods assessed.

That aside, for each instance of the parameter set, the subsequent order of operations is identical. Each run is assigned a new subdirectory, into which all subsequent files are deposited. We calculate GFD and AFD descriptions of the pre-computed conformers according to the trial parameters, and then calculate FDFPs, before calling the benchmarking script. We construct dictionaries mapping from active molecule IDs to targets, conformers, and descriptors, where these can be pre-computed. This is structured so as to minimise the memory footprint of each run, allowing for each set of fingerprints to be computed and loaded only once, although they are involved in multiple comparisons, and to facilitate the easy addition of new descriptors in future.

For the purposes of our study, we consider FDFP variants, ECFP4, ECFP6, USR, USRCAT, and ROCS Shape and Combination modes. For all shape-based approaches, we used the same set of conformers, as described in section 3.2.3. For each molecule, we ordered its set of conformers by their energy under the MMFF94 forcefield, such that the lowest energy conformer is first in the array, with the others following in ascending order, to facilitate the conformer sampling experiment described in section 3.2.3.



0.60	0.75	0.30	0.05	0.10
0.90	0.60	0.45	0.20	0.10
0.60	0.80	0.30	0.30	0.10

0.90	0.80	0.45	0.30	0.10

Figure 3.4: Illustration of the similarity MAX fusion rule, described by Willett et al.[265]. For each column in a $m \times n$ similarity matrix, S, the maximum value is retained in a corresponding column in the n-dimensional vector, $\mathbf{s}$.

ROCS (v. 3.2.2.2) was utilised in two modes, Shape and Combination. The command-line options used were:
'rocs -query -dbase -outputdir -shapeonly -nostructs -progress none -status none -stats all -report each -conflabel title -scdbase -maxconfs 100' and 'rocs -query -dbase -outputdir -nostructs -progress none -status none -stats all -report each -conflabel title -scdbase -maxconfs 100', respectively.

Post-processing of the output text files was with a customised parser (to be found in the benchmarking script file), to construct an equivalent similarity matrix to those obtained with the other methods. USR and USRCAT were calculated for all conformations for each molecule using the default parameters specified in RDKit. ECFP4 and ECFP6-like fingerprints were hashed to a bitstring length of 1024 bits.

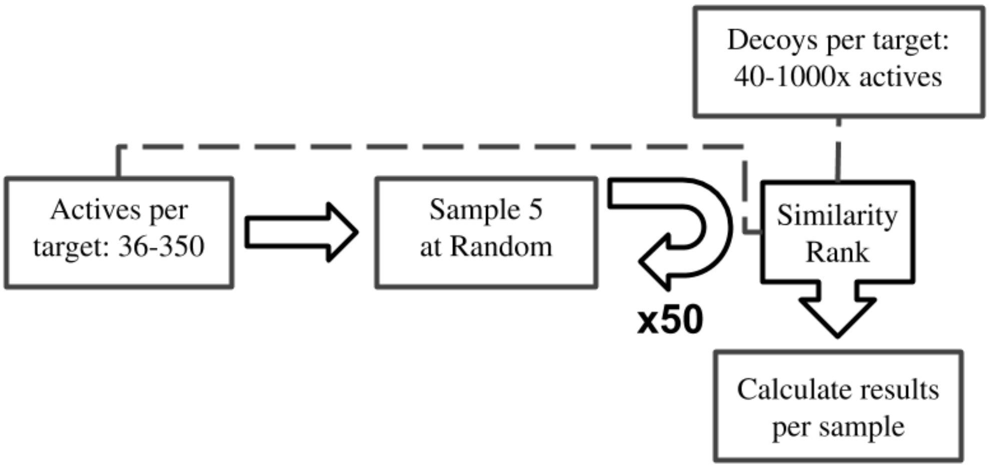


Figure 3.5: Illustration of the multiple-sampling strategy adopted in the benchmarking study. For each target in the three datasets, five known actives are chosen, at random, and without replacement, from the pool of known actives. The similarity of all remaining compounds in the active and decoy sets to these is then calculated, and a max fusion approach taken. This procedure is repeated 50 times in each instance, and results for each sample are calculated and stored for later analysis.

Sampling and screening strategies For each target the same procedure is followed. All compounds, active and decoy, related to an individual target are retrieved, along with their GFD, FDFP, and ECFP descriptors and fingerprints. To assess the impact of search group composition on achieved virtual screening results, we repeat the experiment 50 times for each target, sampling five known actives at random without replacement in each instance. This subset is then used to perform a similarity comparison against all remaining known actives and presumed inactives (as illustrated in Figure 3.5), and performance assessed according to the procedures described in section 3.2.5, Performance metrics. We utilise a group fusion approach, specifically 'max similarity fusion', or max fusion, which has been noted to considerably improve virtual screening performance [265]. For a given panel of five random query molecules, and n database molecules, we condense the resulting $5 \times n$ similarity matrix into an n-dimensional vector, by taking the maximal value for each column. As such, we take the maximum similarity of each of our database molecules to the group of five query molecules, as illustrated in Figure 3.4.

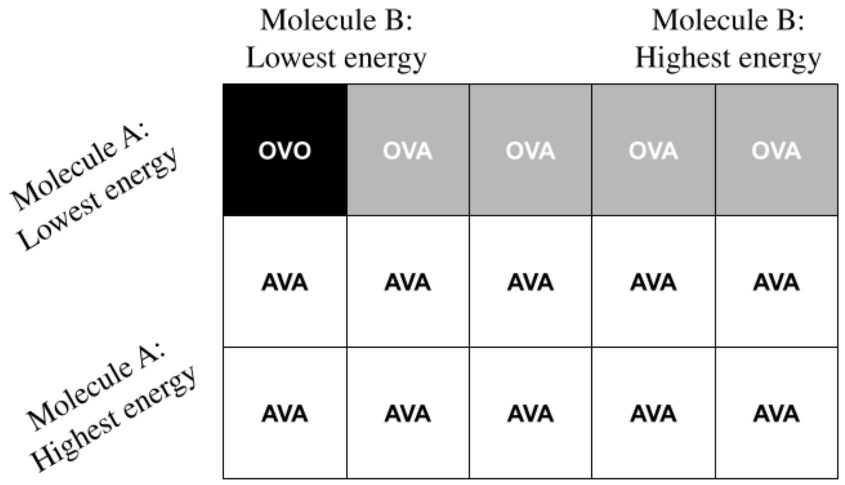


Figure 3.6: Illustration of the three strategies adopted to assess the importance of utilising multiple conformers for shape-based approaches. For two molecules, *A* and *B*, one-vs.-one (OVO), one-vs.all (OVA), and all-vs.all (AVA) describe the result of comparing the lowest-energy conformers of *A* and *B*, the lowest-energy conformer for *A* with all of *B*'s, and all conformers for each molecule respectively. In each case, the result returned is a single, scalar similarity value, obtained by taking the maximum of the cells considered.

In a deviation from typical practice for 2D methods, but one which has frequently been advocated in the literature as a means of improving virtual screening results by reducing the impact of noise from conformer generation [260, 262, 263, 385] we adopted three conformer-set comparison strategies; one-vs.-one (OVO), one-vs.-all (OVA)), and all-vs.-all (AVA), to determine the added value of a multi-conformer approach when conducting shape-based virtual screening. Here, 'one' indicates that the first, i.e. lowest-energy conformer, is used, and 'all' that each conformer is considered. For instance, a one-vs.-all FDFP comparison for two molecules, *A* with three conformers, and *B* with five, would result in a five-element vector, where each position is the similarity of a conformer *B* to the lowest-energy conformer in *A*. In line with the intent of the max fusion approach, we would then take the maximum value of that vector as a representative of the whole. An all-vs.-all comparison would result in a 3×5 matrix, from which we would take the maximum similarity value. This is illustrated in Figure 3.6.

3.2.4 Similarity Ranking

As each method profiled differs slightly, so too does the means of determining their similarity. For FDFP and the ECFP-type methods , we adopt the Dice similarity [202], defined as :

$$S_{\mathrm{A,B}} = 2c/[a+b] \tag{3.1}$$

where $S_{\mathrm{A,B}}$ is the similarity of the binary fingerprint representations of two molecules, A and B, c is the number of 'on', i.e. 1-valued, bits shared between both binary vectors, a the count of on-bits in A, and b the equivalent value for B. These latter quantities are defined as:

$$\begin{aligned} a &= \sum_{j=1}^{j=n} x_{j\mathrm{A}} \\ b &= \sum_{j=1}^{j=n} x_{j\mathrm{B}} \\ c &= \sum_{j=1}^{j=n} x_{j\mathrm{A}} x_{j\mathrm{B}} \end{aligned} \tag{3.2}$$

Dice similarity is faster to calculate than the popular Tanimoto coefficient and results in provably identical similarity rankings [214, 400]. For the ECFP-like methods, we use a bit-vector length of 1024, the default value in the RDKit implementation of that algorithm. Although an increase to 4096 bits was observed to increase performance very slightly in the initial benchmarking study, the additional overhead of doing so means that this is rarely adopted in practice. Longer bitstrings result in fewer bit collisions, and therefore retain somewhat more information, whilst increasing the cost of pairwise comparisons dramatically. For ECFP4 with a bit vector length of 1024, approximately 1.4% of bits set result in collision, having a minimal impact on any subsequent similarity comparisons.

Given that GFD is bounded in [0, 3] for the objects our studies are concerned with, and that more properly that their FD should strictly be in [2,3] assuming an appropriate δ_{max}, we take the Euclidean distance between the GFD of A, g_A and of B, g_B, normalising it with the upper bound, and taking its complement such that the resulting value is 1 for the most-similar pairs [189].

$$S_{\mathrm{A,B}} = 1 - \left(\frac{\sqrt{(g_A - g_B)^2}}{3} \right) \tag{3.3}$$

USR and USRCAT both use the 'USR Score', defined as follows,

$$S_{\mathrm{A,B}} = \left(1 + \frac{1}{N} \sum_{l=1}^{N} |M_l^A - M_l^B| \right)^{-1} \tag{3.4}$$

taking the average of $|M_l^A - M_l^B|$, the Manhattan distance between the value of the USR or USRCAT vectors M^A and M^B at index l, for all pairs of values, which is then translated by unity and inverted, to give a final value which is one for identical vectors, and monotonically decreasing otherwise.

The previous methods have in common that they produce output which is easily contained within a simple vector, and which is not the subject of an optimisation strategy. For the two ROCS approaches, we use the similarity coefficient defined as

$$T_{\mathrm{A,B}} = O_{\mathrm{A,B}} / \left(O_{\mathrm{A,A}} + O_{\mathrm{B,B}} - O_{\mathrm{A,B}} \right) \tag{3.5}$$

where O is the overlap function as defined in Equation 1.1. The ROCS Tanimoto similarity is the overlap in the molecular volumes after alignment, $O_{\mathrm{A,B}}$, divided by the sum of the individual molecular volumes ($O_{\mathrm{A,A}}$, $O_{\mathrm{B,B}}$) minus that overlap. For ROCS combination, this is the combination of shape and colour Tanimoto terms. For simplicity, we halve the resultant value, such that it fits in the same natural range as the other methods profiled.

As previously mentioned, we then adopt a max fusion approach to the similarity matrix generated. The procedure is illustrated in Figure 3.4. More formally, it is defined as the vector, **s** formed from the maximum value of each column in the $m \times n$ pairwise similarity matrix S as follows,

$$\mathbf{s} = \max_{1 \leq j \leq n} S \tag{3.6}$$

where rows correspond to query entries, and columns to library entries. **s** is then ranked in descending order, and the transformation required is applied to the list of identifiers for each molecule in the library. This results in a similarity-ranked list of identifiers, and allows us to determine the relative capacity of each method to separate active and decoy compounds.

3.2.5 Performance metrics

Many schemes for assessing the performance of virtual screening methods have been proposed, utilised, and criticised heavily. To gain a broad overview of the differing characteristics of the methods employed, we adopted many of the metrics utilised in Riniker and Landrum's work, and include some additional analyses based on belief theory [401] to determine whether we can observe any generalisable and statistically-meaningful relationships between similarity and likelihood of activity for the approaches profiled.

The former category contains enrichment-related measures such as the AUC and BEDROC, and diversity-related measures, such as the Scaffold Enrichment Factor (SEF). In accordance with their findings, we limit the number of outcomes here, reporting AUC as an overview statistic, and BEDROC20 as a measure of early enrichment. Early-recognition, or early-enrichment methods, were found to correlate very highly with one another assuming an appropriate choice of parameters. Scaffold-enrichment is also incorporated, as the ability of shape-based approaches to retrieve relatively diverse active compounds, and by doing so to 'scaffold-hop', is a primary motivation for their use and development. In their work, they found that SEF was not especially valuable for 2D approaches, as it correlates very highly with overall performance for those methods, supporting the intuition that such methods are not especially well-suited for 'scaffold-hopping'.

AUC is a common method for assessing overall virtual screening performance, and has the benefits of naturally falling in [0,1], being relatively readily-understood, and providing a global measure of how well a given descriptor-similarity method orders chemical space. Under AUC, a method which ranked a test-set of molecules such that all actives were at the top of the list, followed by all inactives, would achieve a value of 1. Random performance, such that actives are interleaved amongst inactives in accordance with their relative proportion, would result in a value of 0.5. It is defined more formally as

$$AUC = \frac{1}{nN} \sum_{i=2}^{N} A_i \left(I_i - I_{i-1} \right) \tag{3.7}$$

where n represents the number of actives present in a given test set, N is the size of that test set, A the cumulative count of actives, and I the cumulative count of inactives.

BEDROC [216] is based upon a modification of the RIE method, which was developed to avoid some of the difficulties in comparing enrichment factors between different compound sets. To emphasise the importance of early enrichment, RIE uses a decaying exponential weight, as an α parameter, which reduces the contribution of lower-ranked compounds.

$$RIE(\alpha) = \frac{N}{n} \frac{\sum_{i=1}^{n} e^{-\alpha r_i / N}}{\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}} \tag{3.8}$$

where r_i is the rank of a given active, normalised by N, the numerator is the sum of the exponent-weighted normalised ranks of the actives, over the denominator which is representative of the case of their being drawn from a uniform distribution. The result is then transformed with the ratio of the test set size to the number of actives. [216]. We then determine the maximum and minimum RIE achievable with the values for n and N we obtain for each target set, and for the value of α chosen:

$$\begin{aligned} RIE_{\min}(\alpha) &= \frac{N}{n} \frac{1-e^{\alpha n/N}}{1-e^{\alpha}} \\ RIE_{\max}(\alpha) &= \frac{N}{n} \frac{1-e^{-\alpha n/N}}{1-e^{-\alpha}} \end{aligned} \tag{3.9}$$

BEDROC provides a bounded equivalent of the RIE, defined as

$$BEDROC(\alpha) = \frac{RIE(\alpha) - RIE_{\min}(\alpha)}{RIE_{\max}(\alpha) - RIE_{\min}(\alpha)} \tag{3.10}$$

which has the advantage of being comparable between datasets and experiments, and providing a readily interpretable measure in [0, 1]. We adopt BEDROC20 here, as it is the exponentially-weighted analogue of EF(5%), which was found in the original paper to correlate well with SEF below, while avoiding some of the issues associated with EF, especially its instability with varying proportions of actives and inactives.

To calculate scaffold enrichment, we take a similar approach to that used for the calculation of the RIE, without the term for exponential weight decay. We take SEF at 5%, as per the original paper. SEF has the same issues as the regular enrichment factor [212], including a strong dependence on the underlying ratio of diverse actives to inactives, and sensitivity to the percentage value chosen, but is an established and readily-interpreted diversity metric. The definition of a scaffold here is two-fold: we use the BMS [26], and also GMS, which is essentially a pared-back version of the former definition, where information on ring aromaticity and atom identify are removed and replaced with simple heterocycles and carbon placeholders.

SEF can be defined as,

$$SEF(\chi) = \frac{\sum_{i=1}^{n} \delta\left(\min r_i\right)}{\chi^n} \tag{3.11}$$

where $\min r_i$ is the minimum rank for compounds belonging to a given scaffold, χ is the proportion of the ranked compounds to consider, i.e. χ equal to 0.05 entails considering the first 5% of top-ranked compounds, χ^n is the number of unique active scaffolds present in the test set, and $\delta\left(\min r_i\right)$ is

$$\delta\left(\min r_i\right) = \begin{cases} 1, \min r_i \leq \chi N \\ 0, \min r_i > \chi N \end{cases} \tag{3.12}$$

This calculation is identical for both the BMS and GMS definitions of scaffold identity, giving Bemis-Murcko scaffold enrichment (BMSE) and generalised BMS enrichment (GMSE) respectively. The second category of measures, those relating to the construction of probabilistic models which allow us to predict a given compound's activity, given its similarity to the pool of known compounds, is discussed in the following section.

3.2.6 Benchmarking study analysis

Statistical analysis We analysed the distribution of the performance metrics obtained for each target, grouped by database, and by properties of the active compound sets. We applied a ranking procedure, determining the relative performance of each method assessed, under each performance metric implemented, enabling a statistical approach to defining which methods are better at the individual tasks described, and which methods have best overall performance. Our analysis proceeds as per the original paper [214], where they conducted an initial ANOVA test to determine whether there is any statistically significant difference between the methods overall, and then pairwise bootstrapped Friedman post-hoc tests in the event that such a difference is found. ANOVA and Friedman bootstrapped post-hoc tests are performed in the R (v. 3.5.2) statistical computing environment. All other statistical tests, and correlation analysis, are performed using the SciPy (v.1.2.0) statistical package, available in Python3.7. The R script for statistical analysis is provided in the RDKit Github repository (bit.ly/rdkitbenchmark). This script takes as input a series of files (one for each performance metric), which list the methods profiled, and rank of that method for each of the 50-fold repetitions described in section 3.2.3. It outputs a significance table for each method, and an overall significance table. Pre- and post-processing is performed in a Jupyter notebook, taking the results file from each benchmarking experiment, and outputting the formatted significance tables found in Table 4.1.

In addition, for each target, and for each query block, we retrieve the similarity under each method to each remaining active in the test pool, as well as storing the entries used as templates. This allows for post-hoc correlation analysis, to determine the extent to which the various similarity methods produce similar orderings of chemical space, and to what extent that depends on template properties.

Information-theoretic and statistical approaches In addition to the methods employed in Riniker et al.'s work, we take an information-theoretic approach, allowing for consideration of the overall performance of each method, and for the assessment the separability of the distributions. We consider two approaches, commonly employed in the analysis of continuous probabilities, and use normalised similarity distributions as our input. First, we consider the symmetrised form of the Kullback-Leibler divergence (D_{KL}), between the two distributions. This can be thought of as the amount of information that would be lost if we used one distribution to approximate the other, and so captures the degree to which the active distribution, for example, can be modelled by considering the inactive. If the distributions were identical, and there was therefore no information loss upon modelling, the D_{KL} would be zero. This approach has the advantages of a relatively straightforward interpretation and common usage. We take the symmetrised form, as D_{KL} is more properly a measure of the divergence of one distribution from another, rather than the divergence between them, i.e., it is inherently asymmetric. The single form in this case is:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{3.13}$$

with p and q as the probability densities of some distributions P and Q (which we take as the active and inactive distributions, respectively). It can also be defined as

$$D_{\mathrm{KL}}(P \| Q) = \mathrm{H}(P, Q) - \mathrm{H}(P) \tag{3.14}$$

where $\mathrm{H}(P, Q)$ is the cross-entropy of the two distributions, and $\mathrm{H}(P)$ the entropy of P. The asymmetric component here is the lack of consideration of Q, so, to correct for this, we take the sum of $D_{\mathrm{KL}}(P \| Q)$ and $D_{\mathrm{KL}}(Q \| P)$.

As a simplification of the symmetrised form of the D_{KL}, we compute the Jensen-Shannon distance (D_{JS}), which has the advantages of innate symmetry, and fulfilling the prerequisites of a metric, owing to its definition as the sum of the divergences of each distribution with respect the mean. It is defined as

$$D_{\mathrm{JS}} = \sqrt{\frac{D_{KL}(P \| M) + D_{KL}(Q \| M)}{2}} \tag{3.15}$$

where M is the pointwise mean of P and Q.

Finally, we take a statistical approach to assessing the separation of the two distributions, P and Q. We utilise the non-parametric Mann-Whitney U test, which allows us to determine whether two independent samples are drawn from the same distribution. In our case, we employ the test such that we assess whether P is stochastically greater than Q, with continuity correction. If we take both distributions, combine them, and rank their values, while maintaining knowledge of their origin distribution, the U statistic is defined as

$$\mathrm{U} = \mathrm{R} - \frac{n(n+1)}{2} \tag{3.16}$$

where R is the sum of ranks belonging to one distribution, and n the sample size for that distribution. In general, this statistic is calculated for both P and Q, and the smaller of the two used for comparison to the null distribution U statistics. However, as we are interested in the test for P being stochastically greater than Q, i.e. for the $\mathrm{H_A}$ where $\mathrm{P}(P > Q) \neq \mathrm{P}(Q > P)$, we take U of Q only. Results for this test are as p-values, where we take a p-value ≤ 0.05 as sufficient to reject $\mathrm{H_0}$, where $\mathrm{H_0}$ is that $\mathrm{P}(P > Q) = \mathrm{P}(Q > P)$.

FDFP Parameterisation We tested several hundred combinations of parameters, to determine the combination of parameters which produced the best overall ranking. Rank performance of each method was assessed using the pairwise post hoc Friedman test previously described (section 3.2.6). We vary δ_{min}, δ_{int}, and δ_{max} in [0.5, 1.8], [0.2, 0.455], and [3.2, 6, 8, 10], respectively. All units are Å. δ_{min} and δ_{max} parameters were chosen to cover a wide range of values, relevant at the scale of small- and macro-molecules. As such, the upper limits are somewhat higher than were considered in earlier work on this topic [169].

The δ_{min} values take into account the density of vertices placed on the molecular surface ($7/\text{Å}^2$), where the lower of the two is chosen such that $\delta_{min} \geq \sqrt{7}$. For FDFP-specific optimisation, we considered all of the above parameters for calculation of the atomistic FD values, and, in addition, all combinations of fingerprint radii in [2, 3, 4, 5, 6], and bitstring length [256, 512, 1024]. To choose a single-best overall method, we conducted the benchmarking experiment for each combination of parameters, and chose the parameter set which resulted in the best overall ranking.

Probabilistic regression - similarity and activity prediction Given the size and diversity of our retrospective benchmarking set, we utilised these data to construct probabilistic models, providing an evidence-based rationale for choosing similarity thresholds. Our aim is to determine whether we can use such an approach to find support for empirical cut-offs used in making similarity assignations, such as the commonly-used $\geq$ 0.7 Tanimoto similarity threshold used with ECFP4 to assign activity-level similarity, for all methods employed in our version of the benchmarking protocol.

Adopting a similar approach to that described by Muchmore et al.[401], we set out the problem as a probabilistic one, where we wish to define the relationship between similarity to a known template, and probability of a given compound sharing that template compound's activity. In that work, the similarity considered is between pairs of compounds with measured IC_{50} values against a panel of 23 protein targets. They define 'active' compounds as having a measured IC_{50} in the range of one to 10 nM against at least one of these targets, and then randomly sample pairs of actives and inactives from a pool of 60,000 compounds from their corporate database. Dissimilar pairs are those with a differing activity label, meaning that their measured IC_{50} against a given target is greater than 10 nM, that there is greater than 1 log unit difference in their pIC_{50}

at that target, or that they do not have an annotated activity (for the compounds extracted from the corporate database). The distributions of similarity scores are then binned, and the corresponding activities noted, allowing for the fitting of a sigmoidal relationship between the two, in a manner reminiscent of the Hill plot, and for the eventual prediction of shared activity annotations between pairs of compounds held out in a test set.

Differing from their approach, we use the maximum similarity to known actives (max fusion) approach described in section 3.2.3 as our measure of similarity, and utilise the sets of actives and decoys as our labelled training data. During the benchmarking process, we store the distributions of similarity values for each query block to the test set. As well as altering the definition of similarity adopted, we utilise multiple models. First, an isotonic regression model, adapted from its common use in classifier calibration. It is defined as

$$\begin{aligned} &\min \sum_{i=1}^{n} w_i \left(y_i - x_i\right)^2 \\ \text{subject to} \quad & \\ & x_1 \leq x_2 \leq ... \leq x_n, \\ & w \in \mathbb{R}^n \end{aligned} \tag{3.17}$$

such that it minimises the sum of weighted differences, subject to a complete order [402]. It has the advantage of being nowhere-decreasing, which fits better to our intuitive model of the relationship between chemical and biological similarity, whilst also being able to fit sigmoid-like curves with arbitrary precision.

The second approach, logistic regression, is popular for probability-prediction tasks, and is well-calibrated by default. The predicted value, P, is a pseudo-probability of activity, defined as

$$P = \frac{e^{a+w \cdot X}}{1 + e^{a+w \cdot X}} \tag{3.18}$$

where a is a bias term, w the weight corresponding to the independent variable (learned by maximum likelihood estimation) at a given value, and X its value.

We train our model on our decoy and active distributions. Our isotonic regression model is adapted, and logistic regression used as provided, from scikit-learn (v. 0.2.0). We perform 5-fold cross-validation. To assess model quality, we can adopt metrics used commonly in machine learning, such as the Matthews correlation coefficient (MCC), a balanced measure of predictive power, and recall, as a measure of the ability of these models to retrieve active compounds. MCC is defined as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.19}$$

where, in our case, TP is the count of correctly-predicted active compounds, TN correctly-predicted decoy compounds, FP decoy compounds incorrectly predicted as active, and FN the reverse. The MCC gives a balances perspective on overall performance, and is in [-1, 1], with the lower bound indicating perfectly-incorrect classification, and the upper that all predictions are correct. Another use-case, where an end-user is most interested in retrieving active compounds, is assessed through recall, defined as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.20}$$

with all terms as defined for MCC. Initial assessment of model quality is performed on the withheld test set, via these two metrics. Additionally, we validate our models through analysis of endpoints utilising data gathered during the prospective study (section 3.4.2).

3.3 Shape searching for de novo (-)-Englerin A Mimetics

Parts of this section are published as: Shape Similarity by Fractal Dimensionality: An Application in the de novo Design of (-)-Englerin A Mimetics. [403]
Authors: Lukas Friedrich, Ryan Byrne, Michael Mederos y Schnitzler, Aaron Treder, Inderjeet Singh, Christoph Bauer, Thomas Gudermann, Ursula Storch and Gisbert Schneider

GFD Virtual Screening, and SAR To determine the utility of the GFD method in de novo drug design, we applied this approach to identify computationally generated, small molecule mimetics with similar biological activities to the structurally intricate ('complex') natural product (-)-Englerin A, henceforth referred to as compound **1a**. This natural product acts as nanomolar activator of transient receptor potential canonical (TRPC) 4 and 5 (TRPC4/5) calcium-permeable cation channels, which leads to selective growth inhibitions of cancer cell lines, although, interestingly, (-)-Englerin B (**1b**) suffers a complete reduction of activity through the cleavage of the glycolate side-chain [404], demonstrating its importance for the on-target effect.

Utilizing compound **1a** as a template, we previously generated new chemical entities (NCE) by a ligand-based, reaction-driven de novo design method [327, 330]. By topological pharmacophore-based scoring and manual refinement of the computational designs, we identified natural product mimetics inhibiting the transient receptor potential melatstatin (TRPM) (TRPM8) calcium permeable cation channel, also inhibited by compound **1a**. Here we extend this study by utilising the GFD measure. Given that the previously employed design software tool (Design Of Genuine Structures (DOGS)) and the pharmacophore similarity metric (CATS [44, 45]) each rely on two-dimensional molecular representations, we investigated the use of GFD as an orthogonal similarity ranking approach, to take the spatial disposition of molecules into account. By omitting the proposed synthetic routes of the original designs, the library of 903 in silico structures employed in our previous study resulted in a set of 323 unique de novo designed small molecules, as the de novo generation method could employ multiple synthetic routes, resulting in identical molecules.

We ranked these computer-generated designs according to their Euclidean distance from compound **1a** in terms of their GFD. To assess the potential of GFD as a shape-based descriptor for this target case, we conducted a comparative, retrospective, analysis of the chemical space retrieved by this method, against gold-standard fingerprint (ECFP4), moment (USR) and alignment based shape (SHAEP) approaches. Given that we lack a ground-truth in this case, i.e. experimental activity data for each molecule in our compound library, our retrospective analysis adopts two approaches. We begin with an analysis of three data sets; (i) the initial de novo design set, (ii) the 30 top-ranked compounds in terms of global fractal dimensionality distance (GFD distance), and (iii) the top 30 compounds according to their topological pharmacophore similarity (CATS distance) to compound **1a**. Set (iii) is included to compare the GFD ranking approach with the CATS approach described previously. We extracted the BMS of these compounds and analysed their molecular scaffold diversity (pairwise Tanimoto coefficient (Tc)) based on Morgan structural fingerprints (radius = 2), equivalent to ECFP4. This approach allows us to compare the areas of chemical space retrieved by each method, a proxy for the likely on-target efficacy, necessitated by the lack of ground-truth activity data for each library compound. Secondly, we employ an experimentally-validated target-prediction software developed in-house, Self-organizing map–based Prediction of Drug Equivalence Relationships (SPiDER) [405, 406] which utilizes 2D molecular descriptions to provide an estimate of the likelihood of a given compound being active against the target family 'Transient Receptor Potential Ion Channel'.

SHAEP and USR The SHAEP [47] program (v. 1.2.0) was used for comparison with alignment-dependent shape-based methods. The program was run with default settings, in the 'onlyshape' mode (-onlyshape), disabling partial charge contributions (-charge-weighted=False) and limiting output per DOGS molecule to the optimally-overlapping, i.e., best, structure (-nStructures 1). Output was parsed with a Python script (version 3.6.5), taking the 'shape_similarity' field as the output per database molecule, and then taking its complement to obtain a distance measure. The USR descriptors were calculated using the implementation provided in RDKit 2018.03.4 in Python 3.6.5. Scores were calculated using the USR Score implementation using default parameters. We converted these into distance measures by taking the complement of the USR Score.

Scaffold Analysis and Property Calculations Scaffolds were extracted as BMS with RDKit 2017.09 in Python 3.6.5. Physicochemical properties were calculated using RDKit 2017.09 in Python (version 3.6.4). As a prioritisation criteria, we utilised the synthesisability score (synthetic accessibility) [407] implemented in Python (version 3.6.4).

3.3.1 Biological assay

All biological assays were carried out by collaborators at the Ludwig Maximilians University of Munich, or on a fee-for-service basis by Cerep Eurofins. For further detail, see appendix section 5, Bioactivity determination

3.3.2 Analysis of results

We generated pairwise similarity matrices for all compounds based on similarity under GFD and FDFP approaches. FDFP was not used to guide compound selection in this study, it is included to compare the similarity rankings produced by the global (GFD) and atomistic (AFD) approaches. In each case, the resulting pairwise similarity matrix was normalised using a symmetric min-max procedure, defined as

$$minmax(S) = \frac{S - \min S}{\max S - \min S} \tag{3.21}$$

bounding S in [0,1].

3.4 Prospective study

3.4.1 Screening

Seven targets previously utilised in our group for method-benchmarking were selected for analysis. Using an in-house script, we retrieved and standardised activity information from a local instance of ChEMBL 23 for all compounds with high-confidence activity annotations for each receptor in our set. Following this, we impose a threshold, retaining only those compounds with an annotated 'standard value' of ≥ 7, corresponding to a potency of $\leq 100\,\text{nM}$, and with the intended outcome (agonism or antagonism, depending on target, see Table 3.4). This threshold was chosen owing to the ready availability of bioactivity data for these receptors, and with consideration to the concentration at which we intended to test all retrieved compounds (10 µM in the first instance). This represents a 100-fold potency shift between our template compounds and tested retrieved compounds, and was chosen given the abundance of available compounds, the heterogeneity of the assays used to initially profile these compounds, and the known issues of false-positive hits with high-throughput screening approaches [11]. The identifiers for our targets, and the number of template compounds meeting the threshold, are described in table 3.2. Our screening library, comprised of purchasable compounds available from one or more of several suppliers was compiled in-house, and is described on a per-supplier basis in table 3.3.

Our approach here is broadly similar to that described in Section 3.2.3, Sampling and screening strategies, differing in a few crucial aspects. For each target, we sample five known actives from the pool of annotated actives, and use this to calculate a similarity matrix relating the pairwise similarity of our group of query compounds, with all compounds in the remainder of the actives pool and those in the commercial screening library, which we call the test set. The max fusion rule is then applied to condense this into a single similarity value for each compound in the test set, as described in Figure 3.4. For each query set, we note the proportion of the top 20 compounds which are members of the actives set. We then remove these actives from further consideration, and store the identifiers of the top 20 library compounds retrieved, alongside their similarity values. By repeating this process 50 times, we build an incidence table, where we can observe the frequency with which a given compound is highly-ranked with diverse, randomly-chosen query sets. We then choose the twenty most common hits, our 'high-ranked' compounds.

Table 3.2: Target identifiers and number of active compounds retrieved from ChEMBL at the 100 nM threshold.

Short ID	ChEMBL ID	UNIPROT Accession	Recommended Name	Templates at 100 nM
A2a	CHEMBL251	P29274	Adenosine A2a receptor	1,775
CB1	ChEMBL218	P21554	Cannabinoid receptor 1	1,913
GRG	ChEMBL2034	P04150	Glucocorticoid receptor 1	1,315
JNK1	ChEMBL2276	P45983	Mitogen-activated protein kinase 8	508
mGluR5	ChEMBL3227	P41594	Metabotropic glutamate receptor 5	883
PIM1	ChEMBL2147	P11309	Serine/threonine-protein kinase PIM1	2,818
PPAR-δ	ChEMBL3979	Q03181	Peroxisome proliferator-activated receptor delta	538

Table 3.3: Source of screening library for prospective screening campaigns by supplier and collection. Discrepancy in 'total' and 'unique total' owing to availability of the same compound from multiple sources.

Supplier	Compound count
Asinex Elite	104,521
Asinex Fragments	23,179
Asinex Gold + Platinum Collections	296,394
ChemBridge SC Collection	1,032,195
Enamine Advanced	245,930
Enamine HTS Collection	1,678,558
Specs Natural Products	848
Specs Screening Compounds	210,036
Total	3591661
Unique Total	3,380,696

Table 3.4: Biochemical testing carried out on a fee-for-service basis at Cerep Eurofins (Celle l'Evescault, France). Target short-form IDs are matched with corresponding agonist and antagonist assays carried out, with a dash indicating that a given combination was not tested.

Target ID	Agonist Assay ID	Antagonist Assay ID
A2a	4	-
CB1	1744	1745
GRG	469	-
JNK1	-	2880
mGluR5	-	3844
PIM1	-	2919

To assess the overall validity of our approach, we modified the procedure above, to also sample those compounds at around the 1000th rank. Many compounds in this region were also observed to appear frequently, suggesting that the ordering of the library remains relatively stable with the choice of query set. These two pools, of high- and intermediate-rank compounds (*n*=130), were purchased and sent for biochemical testing, which was performed on a fee-for-service basis at Cerep Eurofins (Celle l'Evescault, France) (see Table 3.4 for details of which assays were performed for each target). All compounds were tested at a concentration of 10.0 µM, barring the compounds tested for PPAR-$\alpha/\gamma/\delta$ activity, which were tested by a collaborator (Dr. D. Merk, Goethe-Universität, Frankfurt, Germany) at 30.0 µM, with follow-up EC_{50} determination for those compounds which exhibited activity in the initial screen.

3.4.2 Analysis

Activity data To allow for comparative analysis of all target datasets, compounds tested against PPAR-$\alpha/\gamma/\delta$ with no activity at 30 µM, or with an IC_{50} of $\geq$ 15 µM are treated as inactive. In the larger assay, conducted at Cerep, we define active compounds as those with $\geq$ 30% activation or inhibition of their target, as appropriate, assuming all values below this threshold to be within the noise of the assay, or likely to have an IC_{50} of substantially higher than 10 µM. In the case of CB1, where both agonistic and antagonistic studies were carried out, we merge these values for the overall analysis.

Target prediction We performed post-hoc target prediction for each compound tested using two popular target-prediction programs; the Similarity Ensemble Approach (SEA) and SPiDER. Neither prediction was used when picking compounds. For SEA, we used the public portal provided (sea.bkslab.org), submitted a list of compounds SMILES and specified which target was of interest for each. SPiDER was run locally (Mac Pro, 2012, OS X Yosemite, 10.10.5, 2x2.26 GHz Quad-Core Intel Xeon, 48GB memory) as a KNIME workflow, wrapped in a Python3.7 script. This returns all predicted targets for each compound, and allows for a greater level of detail than the version of that program available online (bit.ly/modlabspider). We tested known active compounds for each target to extract a list of internal target labels of interest, and then parsed the results for our purchased compound to determine which were predicted to be active. Finally, we enriched our initial data with these two predictions as binary labels. For performance assessment, we treat activity as a binary characteristic as discussed above, and binarise predicted pseudo-probabilities at the P(0.05) inflection point. This allows us to compare the performance of these target prediction programs with the target-agnostic similarity-based classifier trained on the benchmarking data, described in section 3.2.6.

3.5 PIM1 - detailed analysis

A novel PIM1 inhibitor was identified in the prospective screening. We searched ChEMBL 25, and PubChem (July 2019) using their online portals, and were unable to find previous evidence of PIM1 inhibition for this compound. Equally, a similarity search for its BMS on ChEMBL at 0.7 Tc with ECFP4, and a substructure search on the PubChem site, did not reveal further relevant records for PIM1. We ordered a kinase panel, also at Cerep Eurofins, to determine the inhibitory efficacy of our compound against other disease-relevant kinases. Additionally, the structure of this compound in complex with the native protein was solved at a fee-for-service provider (SARomics Biostructures AB, Lund, Sweden), to facilitate assessment of the relationship between shape similarity of our compound in its minimised and bound states with other known inhibitors.

3.5.1 Selectivity study

Dataset Construction

While several large studies of kinase specificity have been published [408, 409] we are not aware of a study dataset which covers the group of kinases available as a panel at Cerep. Therefore, we extracted the set of all publicly-known inhibitors available in ChEMBL25 for each of the kinases (*n*=58) included in the panel, using the UniProt IDs collated from Eurofins online material to retrieve compound records using the ChEMBL API [410]. These records were then parsed, to retain only those conducted with 10 µM inhibitor concentration in the presence of 10 µM ATP, emulating our assay conditions.

We identified 167 compounds which had annotated activities for the majority (*n*=55) of the targets. We removed the remaining three targets from further comparative analyses. Of these 167 compounds, 72 had an annotated activity $\geq$ 30% for PIM1 kinase. This exercise was repeated with the PubChem bioassay database, using a Python-based API to retrieve records for assay IDs which were associated with the corresponding Uniprot IDs. However, this approach retrieved 58 compounds which shared a similar number of targets to those retrieved from ChEMBL. 52 of these compounds were present in CheMBL also, and so this dataset was not further considered for the sake of consistency.

Selectivity scores

Several metrics for assessing the selectivity of compounds against classes of targets have been proposed. Simple measures, such as a ratio of actives to inactives at a particular threshold and concentration (S), give an immediate impression of the relative selectivity of compounds. We define this for some threshold, t as

$$S(t) = \frac{\sum_{i}^{n} \theta x_i}{n}, \qquad \text{where } \theta = \begin{cases} 0, & x_i < t \\ 1, & x_i \geq t \end{cases} \tag{3.22}$$

Thresholds at S(50%), S(70%), and S(80%) are reported in the literature [411]. The Gini coefficient has been proposed [412] as a more sensitive, and less threshold-dependent, means of analysing the uneven distribution of activities across a group of targets. It can be defined as [411]

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)x_i}{n \sum_{i=1}^{n} x_i} \tag{3.23}$$

where x are the observed activity values in ascending order and n the number of kinases profiled. Essentially, it is a measure of how the cumulative fraction of the total inhibition (across all kinases) varies with the cumulative fraction of kinases. A compound which is equipotent at all targets would receive a Gini coefficient of 0, and one which had no activity at any target other than that intended a 1. For this ordered case, it is related to the AUC described in equation 3.7 as $AUC = (G + 1)/2$. In its implementation in the paper, and in our reimplementation in Python 3.7, values below 0, or above 100, are capped. Given the assumption, stated above, that inhibition values $0 \leq 30\%$ are within noise limits of our assay system, we set these values to zero. The analysis is also repeated without this cleaning step, for comparison. The authors of that study [412] note the importance of a consistent ATP concentration in assessing inhibition, although they speculate that a linear correction might suffice to correct for this. They observe that this measure is quite dependent on the number of kinases profiled, although a 50-member kinase panel was determined to give good generalisability to the broader kinome. In addition, they state that the ideal concentration range for testing is within 10-100 fold of the compound IC_{50} against the desired target, which renders comparison between compounds somewhat problematic.

Given these issues, and the lack of an established, authoritative, metric in the literature, we adopt several additional selectivity definitions. A recent paper [411] reviews existing measures, and adds two novel types which have similarities to existing, K_D specific measures [409]. We will focus only on those of immediate relevance to the comparison of inhibition percentage values. The first method discussed is the Window Score (WS) a simple proportion of activities within a given percentage of the maximum activity noted for that compound, as follows;

$$\begin{aligned} WS(t) &= \frac{\sum_i^n \theta x_i}{n}, \\ \text{where } \theta &= \begin{cases} 0, & x_i < m \\ 1, & x_i \geq m \end{cases} \\ \text{and } m &= \max(x) - t \end{aligned} \tag{3.24}$$

This approach has the advantage of taking into account the maximal inhibition noted for a given compound, and for considering the relative importance of targets which are significantly inhibited. The authors suggest thresholds for the WS approach at 5, 10, and 20% (WS(5%), WS(10%), WS(20%)). The final method discussed is the Rank Score (RS), the difference between the maximal inhibition noted and that at some specified lower rank. Assuming x is ranked in descending order of inhibition, it is defined as

$$RS(t) = max(x) - x_t \tag{3.25}$$

None of S, WS, RS are comparable between compounds with varying numbers of activities annotated. In all instances where a sensible score cannot be calculated, for example if a compound has no annotated inhibition of greater than 50% and we wish to calculate the selectivity score at that interval, its value is set to be equal to that representing the least specific possible compound, in this case, S(50%) would be set equal to one.

3.5.2 PIM1 - Inhibitor crystal structure

Crystallisation PIM1 apocrystals were soaked with compound **102** at a fee-for-service provider (SARomics Biostructures AB, Lund, Sweden). Final resolution of the solved structure was 1.8 Å. Crystallographic parameters in Appendix 2, Table 2.

Data curation and alignment We downloaded all solved crystal structures for Uniprot P11309/PIM1 (*n*=151) from the RCSB PDB as mmCIF files, employing the programmatic access provided in Biopython. We converted these to PDB files using the pybel (OpenBabel) and Biopython toolkits, removed crystallographic additives, and cleaned them for standard errors using the 'PDBfixer' tool [377], as described in subsection 3.1.3. Sequence-alignment weighted structural-alignment was performed using the Biopython package in Python, following an adapted version of a 'recipe' published online online (bit.ly/weightedalignment). This uses the Needleman-Wunsch global-alignment algorithm [413] of a query crystal protein construct onto the obtained structure to build sequence identity and similarity matrices, and then minimises RMSD with regard to the latter.

Analysis For each PDB file, we extract the ligand in its crystal conformation, generate a set of ten conformers for that ligand (using the approach described in subsection 3.1.4), and store these alongside the apo-protein. For the analysis of ligand crystal-derived and energetically-minimised conformations, we define four sets of pairwise comparisons. 'Self', where we consider the maximum similarity of a given ligand's crystal conformation to those generated by the minimisation routine. 'Crystal-Crystal', where we compare the similarity of the crystal pose of compound **102** to the crystal conformations of all other ligands. 'Crystal-Minimised', which assesses the maximum pairwise similarity of crystal structure of compound **102** against each set of conformers for the remaining molecules. 'Minimised-minimised', which replaces the pose obtained for compound **102** with the set of its generated conformers. Each of these comparisons for the ligand sets is performed using FDFP8 (1024-bit), ROCS Shape, and ROCS combination. We compare binding pockets for each crystal structure in terms of their FDFP8 (512-bit) representation. All FDFP comparisons use the Dice coefficient, ROCS Shape and combination use the ROCS Tanimoto score previously discussed (Equation 3.5). As each method has a different innate scale of similarity values, comparisons are primarily in terms of relative shifts in the distributions, their mean values, and standard deviations. Pairwise statistical analyses use the Kendall-Tau test, at the α=0.05 significance level.

Crystal structure ligand-protein interaction analysis was performed in Schrödinger Maestro 2019 Academic Edition.

3.6 Protein FD - Back Translation and Analysis

3.6.1 Dataset construction

We downloaded the sc-PDB (2017), which contains 17,594 manually-curated ligand-bound crystal structures. Having written a parser to improve record consistency (in some cases, ligands were not explicitly annotated), we were able to reform the structures into a format suitable for automated FDFP generation for 14,556 of these. We then generated a variety of FDFP fingerprints, for all reasonable diameters (FDFP-6,8). These fingerprints were transcribed from array to text format, with the index position of each on-bit (incremented by one) written as a string, for both the target and its corresponding ligand fingerprints, e.g. '1 28 257 0 32 48 100 0'. This representation has the advantage of being pre-tokenised for model building. '0' above represents a period, used to separate the protein and ligand components, and to signal end-of-line.

3.6.2 Models

Frameworks Our proof-of-concept study of the feasibility of translation of a protein pocket FDFP to a corresponding ligand FDFP utilises multiple approaches from the field of sequence transduction. Two different frameworks were employed; OpenNMT, using Torch as a back-end, and the github implementation (bit.ly/aiayn) of the Attention is all you need (AIAYN) transformer, written in Keras, and employing Tensorflow as its backend. The former is used for a survey of existing sequence transduction methods, including 'vanilla' LSTM models, LSTM with attention, GRU, CNN (non-hierarchical), and various formats of the transformer model, based on promising initial results. For an overview of these methods, please consult section 1.5.2. To investigate the effect of hyper-parameter optimisation on model quality and performance, we utilised the AIAYN transformer model.

Models All models were trained on an machine equipped with eight Nvidia GTX 1080Ti GPUs. All models use the same datasets described above. Unless indicated otherwise, batch-sizes are 32, a learning-rate of 0.001 was chosen, the Adam optimiser was utilised for training, dropout is set to 0.1 to slow over-fitting, and a sequence length of 0.4×maximum folded bit-string length enforced. This latter parameter is based on observed bit densities, and excludes less than 1% of pairs at each fold-level. Train test split results in 11,178 training examples, 1,242 test examples.

Models - OpenNMT We assessed the following approaches, utilising the OpenNMT toolkit. Validation was performed every 200 steps. All models were run until a patience callback determined that no improvement had been made within the last 5 epoch-equivalents, at which point model weights were reverted to their then-current state. OpenNMT uses input feeding and global general attention on the decoder by default, so we left this enabled for all models.

1. LSTM: This model was intended as a simple baseline of a naïve sequence transduction approach. The model has 5 layers of LSTM units.
2. GRU: rather than using LSTM recurrent units, this model uses GRU.
3. CNN-3, CNN-9: These models use the CNN2Seq approach, with varying kernel/filter widths corresponding to their titles.

Transformers: All transformer models assessed with OpenNMT used sinusoidal position encoding. Neuron initialisation is by the Glorot approach. Label smoothing loss-function is enabled (0.001), and 4000 warmup steps are hard-coded. Adam $\beta2$ parameter is set to 0.98. Unless otherwise specified, the position-wise feed forward (FF) width is equal to encoder width.

4. MINI-transformer: a relatively shallow and thin transformer, as described in OpenNMT documentation. Model trained with 8 heads. Additional attention types implemented: self (scaled dot), coverage. Adam $\beta2$ parameter - 0.98. Label smoothing enabled, set to 0.001. Gradients computed on each batch (accumulation count=1). Encoder and decoder RNNs 256-wide. 2000 warmup steps.
5. MED-transformer: as per the mini-transformer, although encoder and decoder widths are doubled to 512.
6. MEDWIDE-transformer: as per the MED-transformer, but with 16 heads.

7. OpenNMT-Google: This is the OpenNMT implementation of the AIAYN transformer discussed below. FF width is 2048. 6 layers, accumulation count 2. Stringent gradient normalisation at 0 enabled.
8. OpenNMT-Google Big: essentially the same as OpenNMT-Google, but with 16 heads.
9. OpenNMT-Google Shallow: A two-layer version of the the Open-NMT -Google model.
10. OpenNMT-Google Big-shallow: A shallow (two-layer) version of the OpenNMT-Google Big model.

Models - AIAYN Having observed the relative performance of the models described above, we focused further efforts on optimising the transformer models for our task, and, in addition, assessing the variation in performance associated with small algorithm adjustments between frameworks. All models were run until a patience callback determined that no improvement had been made within the last 5 epoch-equivalents, at which point model weights were reverted to their then-current state. Maximum number of epochs was set to 2000. Maximum length of output sequence set to 160. All combinations of the following parameter sets were assessed:

- Optimisers: Adam, Nadam, Adadelta
- Encoder/decoder width: 256, 512
- Layers (depth) : 2, 4, 6
- Number of heads: 4, 8, 12, 1

We use the 'noam' learning-rate scheduler, as above, with 4000 warmup steps.

3.6.3 Performance assessment

To assess model quality, and to direct further analysis, we take the best average word-wise validation perplexity achieved by each model, as its definition is the same under both frameworks.

This can be defined as

$$
\begin{aligned}
\text{perplexity} &= 2^{-\sum_{i=1}^{N} p(x_i) \cdot log_2 q(x_i)} \\
&= e^{-\sum_{i=1}^{N} p(x_i) \cdot \ln q(x_i)} \\
&= e^{-\sum_{i=1}^{N} \frac{1}{N} \cdot \ln q(x_i)} \\
&= \prod_{i=1}^{N} q(x_i)^{-\frac{1}{N}} \\
&= N \sqrt{\frac{1}{q(x_1) \cdot q(x_2) \ldots \cdot q(x_N)}}
\end{aligned}
\tag{3.26}
$$

where 2 is an arbitrary base, chosen by convention, x_i is the probability of a given word in our corpus, q_i the prediction under our model for that word. From there, taking the natural log for convenience, we assume that each word is equally probable under our base model (random picking), and reach our final definition, which is essentially equal to cross-entropy. Under this definition, then, a uniform model for word probability would give $q(x_i) = \frac{1}{N}$, leaving perplexity equal to N by cancellation. In essence, it is a weighted-average number of choices over the vocabulary, at each position.

To translate this into a metric more familiar and meaningful to computational chemistry, we translate the validation set back into its fingerprint representation (using the beam-search method), and calculate the Tc of the reconstructed ligand fingerprint versus that obtained from the crystal structure. Beam search for translation is a heuristic search algorithm, which, in contrast to greedy-search, keeps k-best intermediate solutions obtained while traversing a subset of paths across the full breadth-first search (BFS) tree of translation possibilities, rather than just picking the option at each step with the largest immediate reward. The best-overall translation is then returned. This is necessary owing to the combinatorial complexity of conducting a full tree search, which has an upper-bound determined by the vocabulary size, raised to the power of the maximum length of the sequence. We set k equal to five. To determine the level of random performance, we sampled an equivalent number of unique on-bits at random from the vocabulary (concatenation of all words) to the count of on-bits for each validation fingerprint, and calculated the Tc between these. This is a relatively harsh benchmark standard, as it draws the correct number of on-bits for each instance, which is not guaranteed for the sequence transduction model. The median bitwise on-bit frequency in the test ligand FDFP dataset is 0.008, with a median of 46 bits set per fingerprint.

3.7 Computational resources

All calculations were run on the ETH Leonhard cluster. Cluster runs CentOS 7.5.1804. CPU-intensive jobs (all where not specified otherwise) were run on mid-range compute nodes equipped with two 18-core Intel Xeon Gold 6140 processors with 384Gb available memory. All jobs were restricted to single compute nodes, as Python cross-node parallelism remains complex. Parallelism was accomplished by logical separation. All CPU-intensive jobs were run with 24 cores and 72 Gb of memory, to allow for rapid scheduling and fair service usage.

GPU-intensive jobs were run on the ETH Leonhard cluster mid-range GPU nodes, which are equipped with two 10-core Xeon E5-2630v4 processors, 256 GB of memory and 8 Nvidia GTX 1080 Ti GPUs. Hyperparameter optimisation was run at full node capacity, all other jobs using 2 GPUs, 6 CPUs for processing purposes, and 128 Gb RAM.

All graphical and statistical analysis was conducted on a high-end laptop, with Intel i7-8750H CPU @ 2.20GHz, 16Gb RAM, Nvidia 1070 GPU. Analyses typically require numpy, scipy, sklearn, matplotlib, and seaborn libraries, and were conducted in the Jupyter lab environment. Package management with Conda. Unless specified otherwise, all statistical analyses were carried out with scipy v1.2.0.

Chapter 4

Results and Discussion

4.1 Benchmarking study

Summary Having developed the FRACTVS package for the shape comparison of small- and macro-molecules, we sought to compare its performance to established 2D and 3D methods on a robust test platform. As previously discussed in the literature [226], similarity methods often have radically different perspectives of chemical space. We adapted the benchmarking approach previously described [214] to facilitate the inclusion of shape-based molecular representations. In addition to the methods described there, we also assessed the performance of shape-based methods with single- and multi-conformer representations of our template and query molecules.

Overall, we found that the FDFP approach developed achieves an equivalent global enrichment to topological approaches when these dominate, and matches ROCS combination performance when trialled on a diverse set. Early-enrichment performance follows a similar trend. Our method has the best overall performance of the methods assessed in the early retrieval of diverse scaffolds. We see some encouraging evidence for the orthogonality of 2D and 3D approaches. Interestingly, we found that our approach varies substantially less than other shape-based methods with the diversity of conformers considered. Additionally, we analysed the distribution of similarity values observed for active and decoy compounds, and trained models to describe the relationship between molecular similarity and biological activity.

4.1.1 Retrospective Virtual Screening Performance

Determining the global 'best' approach to virtual screening depends heavily on the desired outcomes, primarily whether one wishes to promote novelty or early retrieval of active compounds in the similarity-ordered list. In common cases, some combination of these two characteristics is desired, which is reflected in the construction of the benchmarking approach we have adopted [214]. The authors concluded that it was sufficient to include one global enrichment metric, such as AUC, and one measure of early enrichment (such as BEDROC), given the high degree of correlation observed between such metrics. Given that it is commonly proposed that shape-based methods enable users to retrieve more diverse scaffolds, as they are less dependent on explicit atom typing, we include two measures of diversity retrieval. Defining molecular diversity is challenging; here, we use the enrichment in known scaffolds of both the specific (BMS) and generic (GMS) types, where the latter excludes heteroatom typing. Unless stated otherwise, all statistical analyses are performed for the AVA modality, except in the cases of ECFP4 and ECFP6, where this distinction is not applicable.

For each target, from a pool of 88, we repeat an experiment where we take five query molecules, and use them to order a pool of actives and decoys. This is repeated 50 times per target, and the performance of each approach. For each repetition, the AUC, BEDROC20, BMSE, and GMSE are noted for each method profiled. A rank is assigned to each method for each repetition, and the results analysed to give an overall rank for each method on the four tasks.

Results Considering overall performance on the four tasks, in Table 4.1 we see that ECFP6, ECFP4, and FDFP8 are statistically indistinguishable overall, sharing approximately the same averaged rank (3.04, 3.09, 3.19). ROCS combination performs next-best, overall (5.17), and is slightly better than ROCS Shape (5.62), but not to a statistically-significant degree. Similarly, USRCAT (7.15) is slightly, but not significantly better than USR (8.39). The difference between performance for USR and GFD (9.34) is also not statistically significant. However, as can be seen in Figure 4.1, we see variation in the performance of each method over the set of four tasks, and no one method has best performance on the entire set.

For the global enrichment (AUC) task, ECFP4 (2.79), ECFP6 (2.93), FDFP8 (3.10) and ROCS combination (3.74) are statistically indistinguishable. USRCAT (5.28) and ROCS shape (5.33) are in the next bracket, followed

Table 4.1: Results of pairwise post-hoc Friedman test for all similarity searching methods implemented in our version of the benchmarking study. Here, we consider the average rank across all metrics, for each repetition, for each target as our input. 'X' indicates that a given pair has no statistically significant difference in performance, 'o' that the statistic is near to the confidence level, α (0.05), and '-' that there is a statistically significant difference. Similarity searching methods are ordered by rank.

	ECFP4	**FDFP8**	**ROCS Comb**	**ROCS Shape**	**USRCAT**	**USR**	**GFD**	**Rank**
ECFP6	X	X	-	-	-	-	-	1
ECFP4		X	-	-	-	-	-	2
FDFP8			-	-	-	-	-	3
ROCS Comb				X	-	-	-	4
ROCS Shape					-	-	-	5
USRCAT						o	-	6
USR							X	7
GFD								8

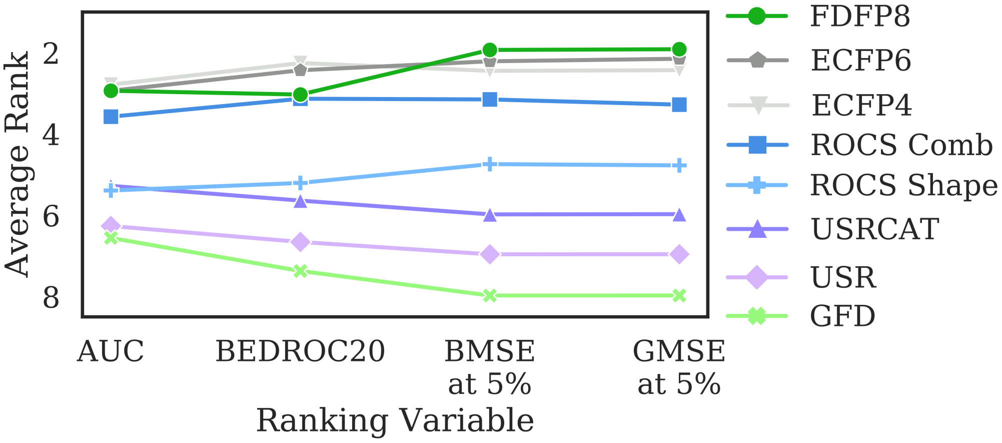


Figure 4.1: Average rank for each task across 50 repetitions for each target. Dark and light grey: ECFP6 and ECFP4. Dark and light green: FDFP8 and GFD. Dark and light blue: ROCS combination and shape. Dark and light purple: USRCAT and USR.

Table 4.2: AUC rank summary: methodology and interpretation as per Table 4.1

	ECFP6	FDFP8	ROCS Comb	USRCAT	ROCS Shape	USR	GFD	Rank
ECFP4	X	X	X	-	-	-	-	1
ECFP6		X	X	-	-	-	-	2
FDFP8			X	-	-	-	-	3
ROCS Comb				-	-	-	-	4
USRCAT					X	-	-	5
ROCS Shape						o	-	6
USR							X	7
GFD								8

Table 4.3: BEDROC20 rank summary.

	ECFP6	FDFP8	ROCS Comb	ROCS Shape	USRCAT	USR	GFD	Rank
ECFP4	X	o	-	-	-	-	-	1
ECFP6		X	-	-	-	-	-	2
FDFP8			X	-	-	-	-	3
ROCS Comb				-	-	-	-	4
ROCS Shape					X	-	-	5
USRCAT						-	-	6
USR							X	7
GFD								8

Table 4.4: BMSE rank summary.

	ECFP6	ECFP4	ROCS Comb	ROCS Shape	USRCAT	USR	GFD	Rank
FDFP8	X	-	-	-	-	-	-	1
ECFP6		X	o	-	-	-	-	2
ECFP4			X	-	-	-	-	3
ROCS Comb				-	-	-	-	4
ROCS Shape					-	-	-	5
USRCAT						-	-	6
USR							-	7
GFD								8

Table 4.5: GMSE rank summary.

	ECFP6	ECFP4	ROCS Comb	ROCS Shape	USRCAT	USR	GFD	Rank
FDFP8	X	-	-	-	-	-	-	1
ECFP6		X	-	-	-	-	-	2
ECFP4			X	-	-	-	-	3
ROCS Comb				-	-	-	-	4
ROCS Shape					-	-	-	5
USRCAT						-	-	6
USR							-	7
GFD								8

by USR (6.26) and GFD (6.55). As can be seen in Figure 4.1, average rank on this task is less variable than for the others, and, as such, it is a less powerful discriminant in this instance.

With the early enrichment (BEDROC20) task, some differences in the rank order are observed. ECFP4 (2.26) and ECFP6 (2.43) form a clear group at the top of the rankings, followed by FDFP8 (3.19) and ROCS combination (3.30, with borderline significance). ROCS shape (5.15) and USRCAT (5.64) are statistically indistinguishable in their performance on this task. Finally, USR (6.66) and GFD (7.39) are statistically indistinguishable.

For the diversity enrichment tasks (BMSE and GMSE), a slightly different pattern is observed. On BMSE, FDFP8 (1.94) and ECFP6 (2.21) are highest-ranked, and statistically indistinguishable. ECFP4 (2.45) and ROCS combination (3.77) are next-best on this task, followed by ROCS shape (4.77), USRCAT (5.97), USR (6.96) and GFD (7.97). With GMSE, the order is preserved, with FDFP8 (1.97), ECFP6 (2.15), ECFP4 (2.43), ROCS Combination (4.77), ROCS Shape (3.77), USRCAT (5.97), USR (6.97), and GFD (7.97).

In summary, USRCAT, USR, and GFD are consistently low-ranked, indicating worse overall performance. ROCS Shape performs moderately well on the AUC and BEDROC20 tasks, doing somewhat better on the BMSE and GMSE 5% tasks, indicating the value of shape-driven approaches in retrieving a diverse subset of the active scaffolds. FDFP8 out-performs all other shape-based methods profiled on the global (AUC) and diversity tasks, and matches ROCS combination on the early enrichment (BEDROC) task (see Table 4.2 and Table 4.3). Both ECFP methods have consistently good performance across all tasks, and are highly correlated with one another. As previously stated, they have statistically-equivalent overall performance to FDFP8 (Table 4.1), but perform somewhat better on the BEDROC20 task, and slightly worse on the diversity tasks.

As this high-level analysis averages performance over the entire set (collating the results for MUV, DUD, and ChEMBL), we thought it informative to consider the per-dataset performance, to determine whether the heterogeneity of these target datasets, on an inter-target and inter-database level, could offer further insight into the relative performance of each method. The per-task, per-database analysis is presented in Figure 4.2. In addition, a comparison of the OVO and AVA modalities is presented, to reflect the importance of multi-conformer ligand representations. Given the similarity in overall performance (see Table 4.1) between ECFP6 and ECFP4, we will primarily refer to ECFP6 henceforth.

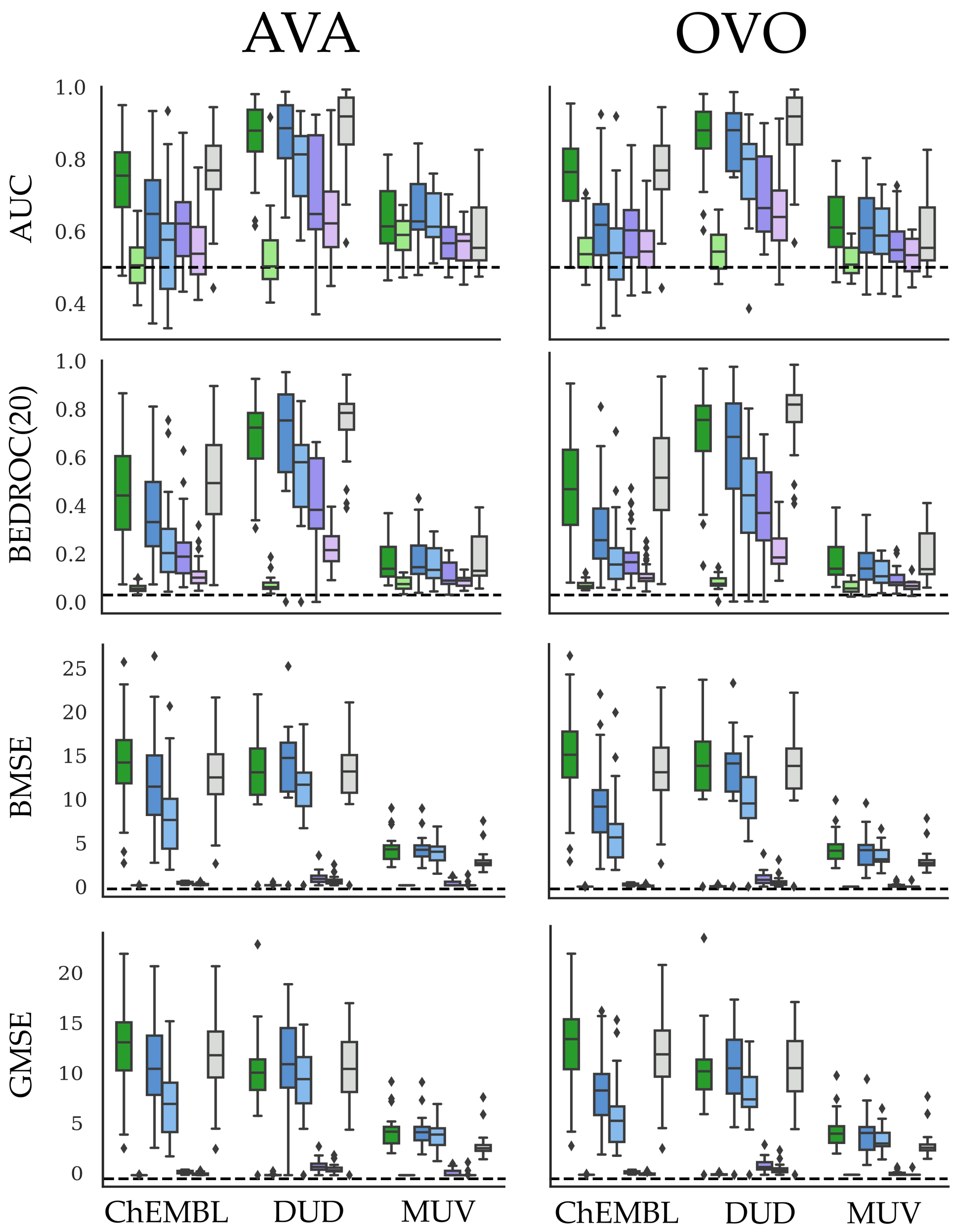


Figure 4.2: Distributions of mean AUC, BEDROC(20), BMSE, and GMSE per target, over 50 repetitions, for the target database (*n*=88), for all-vs.-all (AVA) and one-vs.-one (OVO) configurations. Results for FDFP8 (Green), GFD (light green), ROCS Combination and Shape (Dark blue, sky blue), USRCAT and USR (Dark purple, light purple) and ECFP6 (Grey) are shown. ECFP4 is excluded, for simplicity, but is included in the ranking and statistical analyses.

For the ChEMBL dataset (n=79), we see that FDFP8 has the best AUC and BEDROC(20) results of all shape-based methods tested, and the highest BMSE and GMSE overall. Methods enriched with pharmacophoric information perform well on this set, with ECFP6 having best overall performance, and both enriched shape methods showing improved performance over their purely-geometric counterparts. USRCAT outperforms ROCS Shape on this dataset for both enrichment measures. This relationship is not observed on the scaffold enrichment tasks, where both USR and USRCAT have performance levels only somewhat above random. USR and GFD have notably poorer performance than USRCAT and FDFP, and the worst overall performance on this dataset.

All methods performed well on the DUD targets (n=21) for the global and early enrichment tasks. AUC and BEDROC20 performance for ROCS combination is slightly better than for FDFP8, followed by USRCAT, USR, and GFD. ECFP6 achieved best overall performance on this dataset for these two tasks. For scaffold enrichment, we again see a substantial decline in the relative performance of USR and USRCAT, compared to the other methods profiled. FDFP and ROCS combination have a better performance on these tasks than ECFP6.

FDFP is the second-best method for the MUV (n=17) dataset in terms of enrichment metrics in the AVA configuration, and joint-best in terms of diversity (with ROCS combination performing somewhat better on this task, in the first instance). The relative rankings on this dataset are slightly more complex than in the other cases discussed. For AUC, ROCS shape is the third-best approach, followed by GFD, USR, USRCAT, and ECFP6. For BEDROC20, BMSE, and GMSE, ROCS shape is again third, followed by ECFP6, USRCAT, USR, and GFD.

Overall	AVA FDFP	OVO FDFP	ECFP6	OVA FDFP	AVA ROCS COMB	OVO ROCS COMB	OVO ROCS SHAPE	OVA ROCS COMB	AVA ROCS SHAPE	OVA ROCS SHAPE	AVA USRCAT	OVO USRCAT	OVA USRCAT	AVA USR	OVO USR	OVA USR	AVA GFD	OVO GFD	OVA GFD	Rank
ECFP4	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	1
AVA_FDFP		X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	2
OVO_FDFP			X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	3
ECFP6				X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	4
OVA_FDFP					X	X	X	o	-	-	-	-	-	-	-	-	-	-	-	5
AVA_ROCS_COMB						X	o	-	-	-	-	-	-	-	-	-	-	-	-	6
OVO_ROCS_COMB							X	X	-	-	-	-	-	-	-	-	-	-	-	7
OVO_ROCS_SHAPE								X	X	o	-	-	-	-	-	-	-	-	-	8
OVA_ROCS_COMB									X	X	-	-	-	-	-	-	-	-	-	9
AVA_ROCS_SHAPE										X	X	-	-	-	-	-	-	-	-	10
OVA_ROCS_SHAPE											X	-	-	-	-	-	-	-	-	11
AVA_USRCAT												X	-	-	-	-	-	-	-	12
OVO_USRCAT													X	X	X	-	-	-	-	13
OVA_USRCAT														X	X	X	X	o	-	14
AVA_USR															X	X	X	-	-	15
OVO_USR																X	X	X	-	16
OVA_USR																	X	X	X	17
AVA_GFD																		X	-	18
OVO_GFD																			X	19
OVA_GFD																				20

Table 4.6: Results of pairwise post-hoc Friedman test for all similarity searching methods implemented in our version of the benchmarking study. Here, we consider the average rank across all metrics, for each repetition, for each target as our input. 'X' indicates that a given pair has no statistically significant difference in performance, 'o' that the statistic is near to the confidence level, α (0.05), and '-' that there is a statistically significant difference. Similarity searching methods are ordered by rank. This table includes all modalities - OVO, OVA, AVA.

	ECFP6	OVO	AVA	AVA	OVA	OVA	OVO	AVA	AVA	OVO	OVO	OVA	OVA	OVO	OVO	AVA	OVA	AVA	OVA	Rank
		FDFP	FDFP	ROCS	FDFP	ROCS	ROCS	USRCAT	ROCS	USRCAT	ROCS	USRCAT	ROCS	USR	GFD	USR	USR	GFD	GFD	
				COMB		COMB	COMB		SHAPE		SHAPE		SHAPE							
ECFP4	X	X	X	X	X	X	o	-	-	-	-	-	-	-	-	-	-	-	-	1
ECFP6		X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	2
OVO_FDFP			X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	3
AVA_FDFP				X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	4
AVA_ROCS_COMB					X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	5
OVA_FDFP						X	X	-	-	-	-	-	-	-	-	-	-	-	-	6
OVA_ROCS_COMB							X	-	-	-	-	-	-	-	-	-	-	-	-	7
OVO_ROCS_COMB								o	-	-	-	-	-	-	-	-	-	-	-	8
AVA_USRCAT									X	X	X	X	X	o	-	-	-	-	-	9
AVA_ROCS_SHAPE										X	X	X	X	X	X	o	-	-	-	10
OVO_USRCAT											X	X	X	X	o	o	-	-	-	11
OVO_ROCS_SHAPE												X	X	X	X	o	-	-	-	12
OVA_USRCAT													X	X	X	o	o	-	-	13
OVA_ROCS_SHAPE														X	X	X	o	-	-	14
OVO_USR															X	X	X	X	X	15
OVO_GFD																X	X	X	X	16
AVA_USR																	X	X	X	17
OVA_USR																		X	X	18
AVA_GFD																			X	19
OVA_GFD																				20

Table 4.7: As per Table 4.6, considering only the AUC metric.

	ECFP6	AVA	OVO	AVA	OVA	OVA	OVO	AVA	OVA	AVA	OVO	OVA	OVO	AVA	OVO	OVA	OVO	OVA	AVA	Rank
		ROCS	FDFP	FDFP	FDFP	ROCS	ROCS	ROCS	ROCS	USRCAT	ROCS	USRCAT	USRCAT	USR	USR	USR	GFD	GFD	GFD	
		COMB				COMB	COMB	SHAPE	SHAPE		SHAPE									
ECFP4	X	X	X	o	o	X	-	-	-	-	-	-	-	-	-	-	-	-	-	1
ECFP6		X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	2
AVA_ROCS_COMB			X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	3
OVO_FDFP				X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	4
AVA_FDFP					X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	5
OVA_FDFP						X	X	-	-	-	-	-	-	-	-	-	-	-	-	6
OVA_ROCS_COMB							X	-	-	-	-	-	-	-	-	-	-	-	-	7
OVO_ROCS_COMB								X	-	-	-	-	-	-	-	-	-	-	-	8
AVA_ROCS_SHAPE									X	X	X	o	o	-	-	-	-	-	-	9
OVA_ROCS_SHAPE										X	X	X	X	-	-	-	-	-	-	10
AVA_USRCAT											X	X	X	-	-	-	-	-	-	11
OVO_ROCS_SHAPE												X	X	o	-	-	-	-	-	12
OVA_USRCAT													X	o	-	-	-	-	-	13
OVO_USRCAT														X	o	-	-	-	-	14
AVA_USR															X	X	X	o	o	15
OVO_USR																X	X	X	X	16
OVA_USR																	X	X	X	17
OVO_GFD																		X	X	18
OVA_GFD																			X	19
AVA_GFD																				20

Table 4.8: As per Table 4.6, considering only the BEDROC20 metric.

	OVO FDFP	OVA FDFP	ECFP6	ECFP4	OVO ROCS SHAPE	OVO ROCS COMB	AVA ROCS COMB	AVA ROCS SHAPE	OVA ROCS SHAPE	OVA ROCS COMB	AVA USRCAT	AVA USR	OVO USRCAT	OVO USR	AVA GFD	OVA USRCAT	OVA USR	OVO GFD	OVA GFD	Rank
AVA_FDFP	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
OVO_FDFP		X	X	X	X	o	-	-	-	-	-	-	-	-	-	-	-	-	-	2
OVA_FDFP			X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	3
ECFP6				X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	4
ECFP4					X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	5
OVO_ROCS_SHAPE						X	X	-	-	-	-	-	-	-	-	-	-	-	-	6
OVO_ROCS_COMB							X	X	-	-	-	-	-	-	-	-	-	-	-	7
AVA_ROCS_COMB								X	X	-	-	-	-	-	-	-	-	-	-	8
AVA_ROCS_SHAPE									X	X	-	-	-	-	-	-	-	-	-	9
OVA_ROCS_SHAPE										X	X	-	-	-	-	-	-	-	-	10
OVA_ROCS_COMB											X	-	-	-	-	-	-	-	-	11
AVA_USRCAT												X	X	-	-	-	-	-	-	12
AVA_USR													X	X	X	o	-	-	-	13
OVO_USRCAT														X	X	o	-	-	-	14
OVO_USR															X	X	X	-	-	15
AVA_GFD																X	X	-	-	16
OVA_USRCAT																	X	X	-	17
OVA_USR																		X	-	18
OVO_GFD																			X	19
OVA_GFD																				20

Table 4.9: As per Table 4.6, considering only the BMSE metric.

	OVO FDFP	OVA FDFP	ECFP6	ECFP4	OVO ROCS SHAPE	OVO ROCS COMB	AVA ROCS COMB	AVA ROCS SHAPE	OVA ROCS SHAPE	OVA ROCS COMB	AVA USRCAT	AVA USR	OVO USRCAT	OVO USR	AVA GFD	OVA USRCAT	OVA USR	OVO GFD	OVA GFD	Rank
AVA_FDFP	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
OVO_FDFP		X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
OVA_FDFP			X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	3
ECFP6				X	X	X	o	-	-	-	-	-	-	-	-	-	-	-	-	4
ECFP4					X	X	o	-	-	-	-	-	-	-	-	-	-	-	-	5
OVO_ROCS_SHAPE						X	X	-	-	-	-	-	-	-	-	-	-	-	-	6
OVO_ROCS_COMB							X	X	-	-	-	-	-	-	-	-	-	-	-	7
AVA_ROCS_COMB								X	X	-	-	-	-	-	-	-	-	-	-	8
AVA_ROCS_SHAPE									X	X	-	-	-	-	-	-	-	-	-	9
OVA_ROCS_SHAPE										X	X	-	-	-	-	-	-	-	-	10
OVA_ROCS_COMB											X	-	-	-	-	-	-	-	-	11
AVA_USRCAT												X	X	-	-	-	-	-	-	12
AVA_USR													X	X	X	o	-	-	-	13
OVO_USRCAT														X	X	o	-	-	-	14
OVO_USR															X	X	X	-	-	15
AVA_GFD																X	X	-	-	16
OVA_USRCAT																	X	X	-	17
OVA_USR																		X	-	18
OVO_GFD																			X	19
OVA_GFD																				20

Table 4.10: As per Table 4.6, considering only the GMSE metric.

We repeated the same analysis as above, while considering all conformer modalities, i.e. OVO, OVA, and AVA, in our ranking study. While this necessarily adds some level of redundancy, the end results are important for optimising shape-based screening protocols. Overall, the methods maintain the same ordering, within the noise-levels of the bootstrapped-ranking (FDFP8 and ECFP6 swap positions). For each method, the ordering of approaches is AVA, then OVO, then OVA. No statistically-significant difference in overall performance was noted between all modalities with FDFP8 and USR. OVA was significantly worse for both ROCS Combination, GFD, and USRCAT than the other modalities. ROCS Shape resulted in a near-α p-value for OVO- and AVA configurations. In line with previous studies [64, 260, 263, 414], we found that there is at best marginal benefit in the addition of multiple conformers per molecule.

The observed trend varies somewhat from task-to-task. For instance, OVO-GFD has an indistinguishable performance level to AVA-ROCS Shape on the AUC task, whilst its AVA-configuration is significantly worse (Table 4.7). OVO-FDFP is also higher-ranked than AVA, although this is not statistically-significant at α=0.05. For the BEDROC20 task (Table 4.8), only the AVA configuration of ROCS combination performed better than FDFP, although again the difference is not significant. However, a significant difference was found with OVO-ROCS Combination. Looking at performance on the two diversity tasks (Table 4.9 and Table 4.10), we see that the OVO-configurations of ROCS Combination, Shape, USRCAT and USR achieve better diversity enrichment than their multi-conformer equivalents, and that these differences are statistically significant.

Discussion As discussed by Riniker and Landrum, there is a high-degree of correlation, in general, between all four tasks described, for the simple reason that a descriptor which perfectly ranked the compounds would have maximal AUC and BEDROC20, and scaffold enrichment values, assuming that the percentage chosen for the latter suffices to include all active molecules. Still, given that no method is perfect, and that each gives a different perspective on chemical space [226], it is informative to consider the variation in these other metrics. Diversity-retrieval results obtained for the shape-based approaches are promising. The highest average scaffold enrichment was observed for FDFP, followed by ROCS combination, and then the ECFP methods. While there is a broad similarity between the rankings according to early enrichment and scaffold enrichment, we do not observe a recapitulation of the simple linear relationship observed in the purely 2D study initially conducted (Figure 4.1). For example, although FDFP and ROCS combination have similar and worse early enrichment, respectively, than ECFP6, they show improved and equal scaffold enrichment. This relationship was not noted in the original study by Riniker and Landrum, lending some credence to the assumption that shape-based methods differ substantially from their 2D brethren.

In terms of global and early enrichment, the ECFP methods were noted to have the best overall performance in the initial study, and to be useful tools for scaffold-hopping [415]. Based on a statistical comparison (post-hoc pairwise Friedman test), ECFP6, ECFP4 and FDFP8 are not distinguishable at the α=0.05 significance level, meaning that each approach is approximately equally well-performing when all tasks are considered (Table 4.1). As such, FDFP8, which contains no explicit atom-typing information, is capable of matching the performance of a gold-standard 2D approach, on sets originally constructed using 2D methods. Notably, the developed method demonstrates an improved overall performance over the other 3D approaches profiled, especially in the retrieval of diverse scaffolds. While overall scaffold enrichment for FDFP8 was not significantly improved over an existing fingerprint method (ECFP6), we saw considerable variation in this relationship when considering the results on a per-dataset basis.

The importance of the weighting between pharmacophoric and geometric features seems to differ significantly between datasets, likely reflecting the biases inherent in their construction. For ChEMBL-AUC, we find that FDFP8 performs well, but that other shape-based approaches have significantly poorer enrichment than an established 2D method. This relationship is also seen with the early enrichment task, albeit to a lesser extent. For MUV-AUC, however, all shape-based methods, FDFP8 included, out-perform ECFP6. One might infer that this approach, and other purely shape-based methods, capture pharmacophoric type information implicitly, but this is hard to quantify given that there is no purely-geometric method with which to compare it. A useful proxy, in our case, is to compare the performance of the 'unenriched' and 'enriched' shape-based methods, USR and ROCS Shape on one hand, and USRCAT and ROCS combination on the other. This makes the simplistic assumption that the unenriched approaches do not innately encode pharmacophoric information. The source of better overall performance with these 'enriched' methods is not entirely clear, although the intuitive explanation is that it captures additional information necessary for binding. For the ChEMBL dataset, we see a substantial improvement in each metric for each of the enriched methods over their unenriched counterpart. This relationship is preserved for DUD, but to a lesser extent for MUV, where the relationship is more ambiguous. Given this, it seems reasonable that the variation is owing to the inclusion of the pharmacophoric information; equally, however, it could be that these methods simply down-weight the importance of a noisy shape-component.

However, as the MUV dataset is the only one which sees improved performance overall for shape-based approaches than 2D comparators, and as the enriched methods offer little additional benefit in this instance, it suggests that the former hypothesis may have more support from these data. The marginal value of pharmacophoric information seems limited here, likely owing to the efforts made by the original authors of the MUV set to ensure that active compounds were embedded within decoy compounds on a topological basis. The performance of ROCS shape and combination modalities here are in line with an earlier study on the same dataset [396]. As the active to decoy ratio is at least five-times larger than for the other datasets considered, it is difficult to ascertain whether this variation is fundamental to the sampling strategy employed in its construction or not. In contrast with an earlier study which focused solely on the DUD dataset [416], we see a more complex overall picture for the relative performance of USR-type approximate methods, and ROCS-type alignment-based approaches, when we consider the other two datasets.

While the alignment-based approaches demonstrate better retrieval performance, there are a few indicators that the moments-based approaches offer a novel perspective on chemical space, such as their performance when compared to ECFP6 on the MUV-AUC experiment, and solid outcomes for the ChEMBL-AUC experiment as well. USRCAT shows markedly better overall performance than its equivalent without pharmacophoric information, achieving better average AUC than ROCS shape on the ChEMBL set, and remains inexpensive to calculate. While the results are not quite as good as those observed with ROCS, there is a place for such fast, approximate methods, especially in the consideration of large chemical spaces. Given that this does not correlate with performance on the ChEMBL-BEDROC(20) task, it seems plausible that ECFP6 picks what few ‘low-hanging fruit’ are remaining in the somewhat sparse MUV landscape, as reflected in the limited diversity of the retrieved compounds (see BMSE and GMSE performance).

Somewhat surprisingly, overall performance on this dataset was not worse than on ChEMBL, and in some cases significantly higher, although this was certainly not the case for the purely topological approach ECFP6 which saw a substantial degradation in performance on all tasks barring BEDROC20, as previously discussed. One explanation for this observation is that, for any given query molecule, compounds which possess a corresponding electronic disposition and a matching shape are necessarily a subset of those with matching molecular shapes. By adding this information as an additional ordering on the space of compounds, it appears that such methods substantially improve early enrichment. As such, it seems plausible that one benefit of pharmacophoric methods resides in their exclusion of dissimilar compounds, which can have similar shapes given a broad enough sampling of conformational space. Given that MUV set is constitutionally difficult to separate based on simple properties such as charge, the smaller difference observed in global performance between enriched and unenriched methods for that set would seem to fit that hypothesis. Additionally, the improvement in AUC for ROCS shape, USR, and GFD when going from ChEMBL to MUV has, to the best of our knowledge, not been observed for any of the 2D methods profiled on this platform previously. Again, this suggests a certain orthogonality between the information captured by geometric and 2D approaches, given that different sets are more separable according to one family of similarity approaches than another.

Returning to the DUD dataset, the results obtained support the view, in line with comments made by the creator of that resource [236], that it is largely inappropriate for distinguishing between LBVS approaches. Previous work comparing shape-based approaches using this set [416] concluded that USR and ROCS shape were approximately equally powerful discriminants for DUD actives and decoys, based on AUC alone. While we see similar AUC values for USR to those published, our AUC for ROCS shape is substantially higher (approximately 0.19) higher than found by the authors. Our value for ROCS combination AUC is also somewhat higher, 0.92 as opposed to to approximately 0.75 in that study. This likely reflects our overall screening methodology, utilising the max fusion approach for example, or some improvement in the ROCS codebase. It is interesting that we see no substantial increase in USR AUC in this instance, however. USR and USRCAT achieved reasonable performance on this dataset, indicating that they capture sufficient information to distinguish between topologically-distinct molecules.

We see relatively poor scaffold enrichment for USR and USRCAT for this dataset, only somewhat indistinguishable from random performance. While this bears further investigation, a plausible interpretation is that this reflects both the relatively low scaffold diversity of the DUD set (0.47 and 0.32 for BMS and GMS respectively), and the higher ratio of actives to inactives (approximately an order of magnitude higher than that for MUV), imposing limits on achievable scaffold enrichment beyond random. Given the homogeneity of many of the DUD active sets, it is possible that these approaches preferentially retrieve one or two chemotypes, explaining their decent overall performance. However, given that the previous study concluded that a simple heavy atom count sufficed to give equivalent performance to both USR and ROCS shape, the utility of such comparisons is unclear. This highlights an important issue with the approach adopted, in that the ranking procedure substantially overweights one dataset (ChEMBL) by virtue of its higher proportion of the targets assessed. Equally, the arbitrarily-good performance of LBVS approaches on the DUD dataset renders comparison somewhat ineffectual, and calls into question the merit of merging these ranks to the overall ranking.

Overall, ROCS combination is the 'winner' for the DUD dataset amongst the shape-based approaches, likely reflecting its incorporation of pharmacophoric features in the colour score, closely followed by FDFP. As such, and given the composition of the DUD active and decoy sets, it is likely in this instance that pharmacophoric features are a much better discriminator than geometric ones. This does, however, again raise the question of what we mean by shape; GFD and USR are purely shape-based approaches, but have approximately random enrichment here, whereas an alignment-based approach sees considerable improvement from baseline. Whether this is a trivial matter, owing to the relative ease of alignment of two topologically-similar molecules, is uncertain, and requires further investigation. Regardless, the conformational space sampled, and topological representation of same, suffices to render FDFP approximately equivalent to ROCS combination and ECFP6 in this instance, lending support to our conclusion that FDFP captures pharmacophoric information through its combination of geometry and topography.

In Figure 4.2 we see some interesting relationships, which might help us test that understanding. USR, for example, shows a decline in performance for the AUC task (ChEMBL and DUD sets) while going from OVO to AVA. However, the opposite is true for MUV, where there is a substantial increase, with a higher average AUC than USR or USR-CAT. The added benefit of multi-conformer representations is clearer, overall, for MUV than the other sets, for all shape-based methods other than FDFP. Why multi-conformer representations should 'strengthen' the signal for MUV, and add noise for the other methods is unclear, but we observe that the 'unenriched' methods have substantially higher variability with AVA than OVO configuration. Despite its surprisingly competitive performance in the MUV-AUC comparison, GFD seems too coarse-grained a method for blind-use against large sets. USR, which has a similar overall performance to GFD, has demonstrated utility in prospective studies for arylamine N-acetyltransferases[88], and, as such, it is likely that the utility of these coarser methods depends to a large extent on the properties of the template ligands and screening databases in question.

In general, the addition of multiple conformer representations per query and target molecule offers at best a moderate boost to enrichment and diversity metrics for all shape-based methods. We observed an approximately equal increase in power with the OVA and AVA configurations, although our experimental setup does not let us deconvolute these results to assess the per-query molecule shift, so it is possible that one or other configuration has some marginal benefit, and that this is lost in the double max-fusion synopsis. These results are somewhat counter-intuitive, as one might imagine that the benefits of conformational diversity observed with the AVA case should be preserved for OVA; while OVA is statistically-indistinguishable for FDFP, USR, GFD, and ROCS Shape, entirely shape-based approaches, this is not maintained for those methods which incorporate pharmacophoric information. In addition, improved performance in diversity sampling with lower conformer numbers is a surprising outcome, but one which may be valuable for future large-scale screening efforts.

As with previous works in this field [416, 417], we observed no significant correlation between the performance of the methods described and the number of rotatable bonds in the query compounds. In addition, we found no significant correlation between AUC per iteration and the average synthetic accessibility (a proxy for complexity) or molecular weight for any of the methods profiled. Previous work has found that heavy atom count has some influence on the achieved enrichment [416], so it is unclear whether this disparity is owing to some 'averaging out' effect of the max fusion approach, a property of the conformer generation strategy [390, 418], or a result of the procedures employed in dataset construction [419]. We did not use the proprietary OpenEye OMEGA[390, 418] conformer generation software, instead preferring an open-source approach, to facilitate the analysis and dissemination of our software. By default, the OMEGA approach utilises a modified form of the MMFF94 forcefield, ignoring Coulomb interactions[420], and the attractive component of van der Waals[390]. As such, we feel the RDKit MMFF94 implementation to be a reasonable substitute, and a fair method for all trialled software, given that application of the full forcefield to OMEGA conformations has been observed to improve ROCS performance [385].

Overall, FDFP8 behaves like a shape-descriptor, with considerably better performance on the MUV set than purely topological methods, while maintaining best-in-class performance on the ChEMBL dataset. Its relative invariance with number of conformers chosen could be an artefact of the conformer generation approach chosen (perhaps a more substantial change would be seen with a larger conformational range), but indicates that this approach could be used in a single-conformer mode without a major decline in performance. Whilst somewhat surprising on its face, this is in line with previous findings, that found multi-conformer representations to be of limited value [64, 219, 260, 414], certainly in comparison to the multi-query approach [263]. How to choose an ideal set for multi-query searching is somewhat of an open question, although diversity-focused methods have been noted to substantially improve performance. Even simplistic multi-query approaches such as that trialled here resulted in substantial gains in enrichment [396].

4.1.2 Distribution and probabilistic analysis

While AUC and BEDROC20 are useful single-number representations of similarity distributions, one can also take a more general approach to their analysis. In principle, an ideal similarity metric should maximise the Jensen-Shannon divergence (D_{JS}) between P and Q, where these are the distributions of similarity values for the active and inactive set, respectively. This formalism allows for a numerical analysis of the distribution graphs shown in Figure 4.4, in table 4.11.

Results Taking the performance of FDFP8 as an example, we see substantial differences in P and Q, the similarity distributions for the active and inactive sets, for the three datasets. In Figure 4.3, DUD shows a multimodal distribution for P, with a substantial rightwards shift. ChEMBL has next-most separation between distributions, followed by MUV. The results suggest that the majority of the DUD actives are not embedded

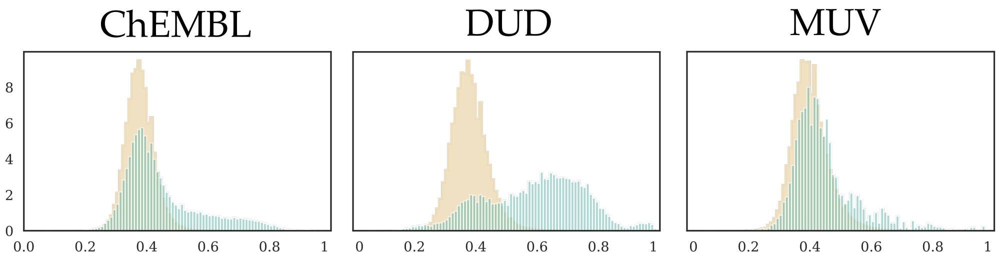


Figure 4.3: Distribution of similarity values for actives (green) and decoys (orange) on a per-dataset basis for the FDFP8 descriptor. The relative difficulty of the overall and early enrichment metrics are a function of these distributions. ChEMBL and MUV active distributions resemble that o the decoys, but with more positive skewness.The multimodal distribution obtained for DUD is qualitatively different.

Table 4.11: Jensen-Shannon divergence (D_{JS}) for the distributions of active and decoy similarities retained from the benchmarking experiment. D_{JS} and rank (in parentheses) are provided for the distributions in Figure 4.4, and for the distributions for each dataset individually. It was not possible to calculate a sensible D_{JS} for GFD. Text in grey indicates those combinations for which the null hypothesis, that the similarity for actives is not stochastically greater than that for decoys, is not rejected at the 0.05 significance level, under the Mann-Whitney U test.

	Overall	**ChEMBL**	**DUD**	**MUV**
ECFP6	0.36 (1)	0.42 (1)	0.60 (1)	0.23 (1)
ECFP4	0.36 (1)	0.42 (1)	0.59 (2)	0.21 (1)
FDFP8	0.34 (3)	0.40 (3)	0.58 (3)	0.20 (3)
ROCS Comb	0.33 (4)	0.36 (4)	0.58 (3)	0.20 (3)
ROCS Shape	0.25 (5)	0.26 (5)	0.45 (5)	0.19 (5)
USRCAT	0.17 (6)	0.21 (6)	0.36 (6)	0.13 (6)
USR	0.09 (7)	0.10 (7)	0.20 (7)	0.09 (7)

within their decoy set, as far as this approach is concerned, which is also reflected in the computed D_{JS} (Table 4.11).

Considering overall separability of P and Q for the combined datasets, we see that that there is a clear overall relationship between the degree of separability of these distributions for each method, and the achievable enrichment factors. This is apparent from inspection of Figure 4.4, and supported by Table 4.11. Ranking methods based on D_{JS} results for the

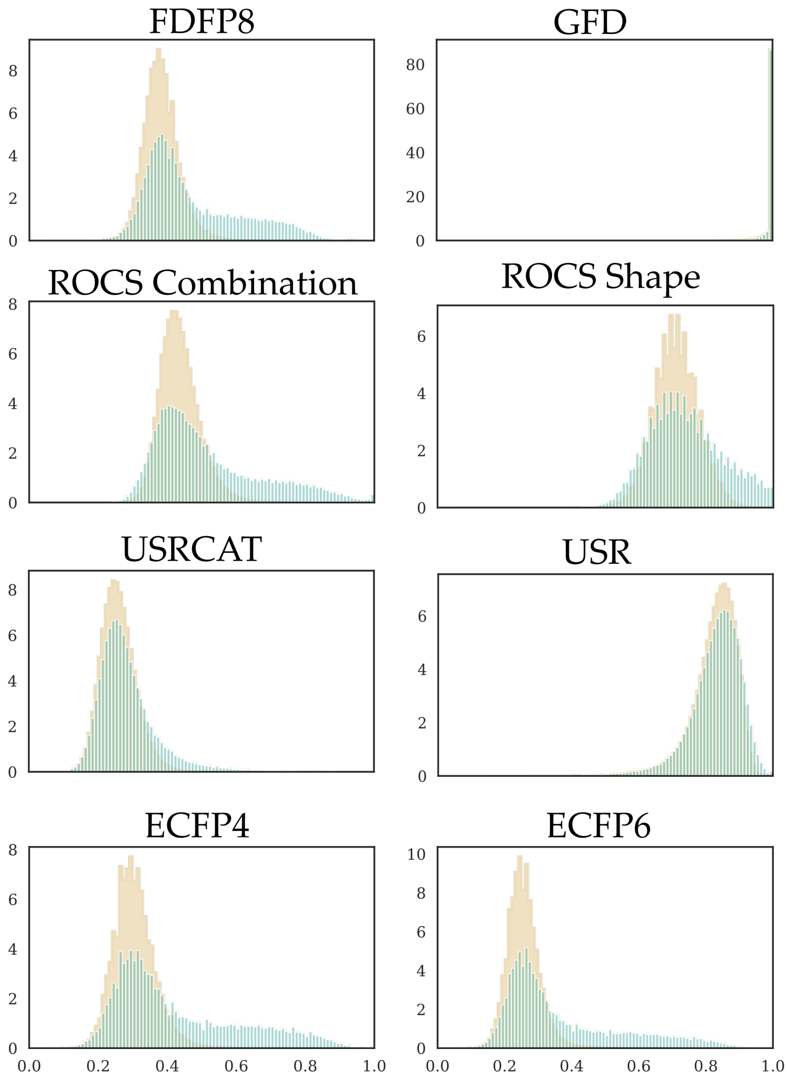


Figure 4.4: Distribution of similarity values for actives (green) and decoys (orange) from the combined results obtained from the benchmarking process. In the cases of FDFP8, ROCS Combination, Shape, and ECFP approaches, we see a clear separation in the distributions, with a heavy-shoulder on the actives distribution. USR and USRCAT have much less obvious separation of actives and decoys. Given the extremely narrow range for GFD, it is difficult to estimate performance from this graph. (section 3.2.2)

Table 4.12: Mean and standard deviations for the active (P) and decoy (Q) distributions, by method and dataset. These distributions are not normally-distributed (normality test, SciPy [421, 422], returns a p-value below the 0.05 threshold in all instances, recommending rejection of the null hypothesis that each distribution is normally-distributed). Methods are ranked in order of performance on the overall-AUC task.

Method	Dataset	$\overline{P}$ ($s\overline{P}$)	$\overline{Q}$ ($s\overline{Q}$)
ECFP6	DUD	0.50 (0.19)	0.25 (0.05)
ECFP6	MUV	0.32 (0.13)	0.27 (0.06)
ECFP6	ChEMBL	0.39 (0.18)	0.25 (0.04)
ECFP6	Overall	0.37 (0.18)	0.25 (0.05)
ECFP4	DUD	0.57 (0.19)	0.29 (0.07)
ECFP4	MUV	0.37 (0.13)	0.32 (0.07)
ECFP4	ChEMBL	0.44 (0.18)	0.30 (0.05)
ECFP4	Overall	0.43 (0.18)	0.30 (0.06)
FDFP8	DUD	0.72 (0.11)	0.60 (0.07)
FDFP8	MUV	0.66 (0.08)	0.62 (0.07)
FDFP8	ChEMBL	0.64 (0.09)	0.58 (0.06)
FDFP8	Overall	0.64 (0.10)	0.59 (0.06)
ROCS Comb	DUD	0.68 (0.16)	0.45 (0.06)
ROCS Comb	MUV	0.53 (0.12)	0.47 (0.06)
ROCS Comb	ChEMBL	0.49 (0.15)	0.41 (0.05)
ROCS Comb	Overall	0.51 (0.16)	0.43 (0.06)
ROCS Shape	DUD	0.84 (0.10)	0.74 (0.06)
ROCS Shape	MUV	0.78 (0.08)	0.74 (0.07)
ROCS Shape	ChEMBL	0.71 (0.10)	0.69 (0.06)
ROCS Shape	Overall	0.74 (0.11)	0.71 (0.07)
USRCAT	DUD	0.34 (0.13)	0.26 (0.05)
USRCAT	MUV	0.30 (0.07)	0.28 (0.05)
USRCAT	ChEMBL	0.28 (0.08)	0.25 (0.05)
USRCAT	Overall	0.28 (0.09)	0.25 (0.05)
USR	DUD	0.85 (0.07)	0.82 (0.07)
USR	MUV	0.85 (0.06)	0.84 (0.06)
USR	ChEMBL	0.82 (0.08)	0.81 (0.07)
USR	Overall	0.83 (0.08)	0.82 (0.07)

combined distributions results in an identical overall ranking to that achieved with AUC. The approach is generally sufficient to discriminate between methods for each dataset, and provides additional numerical support for the perceived 'difficulty' of each dataset, and another perspective on the AUC values achieved. We do see some deviation from the distributions of average AUC recorded in Figure 4.2, possibly reflecting the variation in P and Q on a per-target basis.

Secondary analysis As a consequence of our modifications to the benchmarking approach, we retained data on similarity distributions under each approach, and the recorded similarity values per active compound per repetition of the experiment. As the discussion of similarity under a new method often ends with the question 'how similar is similar enough?', we utilised these data to construct models to approach answering that question. We assessed two approaches, both of which take the form of a regression, where the outcome is the pseudo-probability of activity. As we utilise the results of our prospective application as a test-set, we will defer further discussion to subsection 4.3.3, for the sake of clarity.

Discussion We see some broad similarities in the similarity distributions between some of the approaches, which can be used to explain the single-valued performance metrics discussed above, and to derive some novel metrics [423]. ECFP4 and ECFP6 are very similar, as might be expected. Both USR and ROCS shape have a strongly right-shifted distribution for both active and decoy sets, indicating that, to a certain extent, many molecules 'look the same' under those approaches. This might serve to explain the observed scaffold-hopping tendency of these approaches, but it does not imply equality. Notably, ROCS shape has a clear separation in the two distributions beyond 0.8 similarity. USR also separates the two at a higher similarity level, but fewer of the compounds are in this region, limiting the achievable global and early enrichment. Interestingly, USR displays a distribution of similarity values where the left-tail, i.e. the more dissimilar compounds, is heavier than the right, which inverts the relation seen in all other distributions here, and the pattern typically observed with virtual screening methods [204]. USRCAT behaves much more like ROCS combination and FDFP than its unenriched version. Separation in the right-tail is considerably better, although still less obvious than for ROCS combination, FDFP, or the ECFP fingerprints. The distributions of these last three are substantially similar, although this obscures relevant parameters such as the relative ranking of diverse scaffolds.

When comparing the achieved AUC and BEDROC20 (Figure 4.2), we find that the D_{JS} correlates well with the ranking observed by AUC, indicating that this information-theoretic measure of distribution separability is useful in comparing methods. Equally, it suggests that analysing these distributions in greater detail might help elicit useful cut-off criteria; this will be discussed in subsection 4.3.3. Interestingly, D_{JS} and MWU results are not in perfect alignment. Although USRCAT sees a smaller separability under JSD than ROCS shape, for each database we reject the H_0 that $P > Q$ at the 0.05 significance level, which is not the case for ROCS shape. For DUD and MUV, there is some disagreement between the two measures; although the D_{JS} values are higher for both datasets under ROCS shape than those seen for USRCAT, we see a failure to reject the H_0 for these datasets only in the case of ROCS shape. When comparing the achieved AUC and BEDROC20 (Figure 4.2), we find that the D_{JS} order better recreates the relative performance of the methods. This may be owing to some assumptions of the MWU test, namely that P and Q should have approximately equal variance. However, in Table 4.12, we see that variability for Q remains relatively constant for every method, while the standard deviation of P increases in line with overall increase in performance on the AUC task. Accordingly, both USRCAT and ROCS shape have comparatively similar $\overline{P}(s\overline{P}$ and $\overline{Q}(s\overline{Q}$ values, rendering it less likely that variation is the cause of the discrepancy, and somewhat more likely that it results from using a median-sensitive statistical test for a problem where extreme-values are most often of interest.

4.1.3 Algorithm Run-time Performance

A frequent criticism of 3D methods in virtual screening is the added overhead compared to lower-dimensional approaches. This relatively-heavy computational requirement serves as a barrier to the wider adoption of these methods. As such, methods to reduce this could facilitate broader usage of shape-based approaches, and their integration into the virtual screening pipeline at an earlier stage [424].

Results If we take the similarity benchmarking experiment as an example, generating ECFP6 fingerprints for each dataset, and then performing the repeated similarity screening experiment described in Figure 3.5 takes 35±3 minutes wall-clock time on a 20-core machine (5 repetitions), which includes app. 57 million pairwise similarity comparisons, of which just over one minute is required for fingerprint generation. All times to follow are wall-clock on the machine described in section 3.7, but in each case are trivially parallelisable.

Conformer generation is typically the rate-limiting step for alignment-independent shape-based methods. Sampling 1000 molecules at random from the three databases, conformer generation using the procedure previously described takes 18±8 seconds. For the complete set of molecules in the three databases (n=176,225), conformer generation takes 57 minutes. USR and USRCAT vector description and similarity calculation are then only slightly slower than the ECFP method. Given that approximately 5 conformers are retained per molecule (n=939,550), the influence of this secondary stage is multiplied by approximately 5^2. As such, the benchmarking experiment takes approximately two hours for USR or USRCAT. The primary cause for the discrepancy in observed and expected times is due to the much larger number of pairwise comparisons which are made on each repetition, owing to the number of conformers per molecule, and on the additional post-processing of this data to give a sensible output.

FDFP8 calculation, including molecular surface generation, FD calculation, and the fingerprinting process, takes an average of 35±7 seconds over 50 random samples of 1000 molecules from each of the datasets, or 55±12 if minimisation is included. Minimisation aside, larger molecules take slightly longer, on average, but this effect plateaus at a certain point, with macromolecules taking up to ten times as long as small molecules. Fingerprint generation takes approximately eight minutes. In total, then, the first run using the FDFP8 method takes approximately nine hours.

Subsequent runs take approximately as long as the ECFP method, as the fingerprints can be loaded into memory, or recomputed on the fly, in a reasonable time frame.

The equivalent procedure (combining ROCS combination and ROCS Shape) , takes a total of 95 hours, plus initial conformer generation. The expense primarily comes from the computational cost and complexity of pairwise alignment. In short, ECFP is more than an order-of-magnitude faster than USR/USRCAT when considering conformer generation. In turn, these methods are roughly five times faster than FDFP, which is, in turn five times faster than ROCS for its initial calculation, and more than two orders faster for subsequent repetitions of the experiment.

The overhead of calculating and retaining additional metrics means that it is difficult to generalise from these figures. A simple implementation of pairwise similarity calculation for ten conformers of compound **1a** against the 'immediately available' subset of ZINC15 (*n*=10,285,641) takes 40±2 seconds (5 repetitions) on a single-core machine. The same calculation would take between 50 and 100,000 seconds on an equally powerful CPU, according to data from the OpenEye ROCS website, or 100 seconds using their recent FastROCS GPU-accelerated implementation (bit.ly/FastROCS).

Discussion In practice, this means that shape-based screening against millions of compounds can be accomplished with our method in a similar time-frame to traditional 2D approaches, assuming that the libraries are pre-computed, and on widely-available hardware, such as a commercial-standard laptop. To the best of our knowledge, no publication has been released to date discussing the relative characteristics of ROCS vs. Fast ROCS, so it is unclear to what extent the analysis performed here generalises to results obtained with that approach.

One would imagine that a commensurate speed-up could be accomplished with the initial computation, and subsequent pairwise similarity, under our approach, given that in both cases the rate-limiting process is based on pairwise distance calculation, which is trivially parallelisable. A 20-40 times speed-up on similarity searching [425] and a linearisation of pairwise similarity searches on conventional hardware have recently been demonstrated for other binary vectors [426, 427]. Additionally, the FDFP approach is property-independent in the comparison stage, meaning that highly-flexible molecules impose no additional overhead, for example. These rates of comparison allow for the shape-based consideration of a much larger swathe of chemical space than has previously been feasible in a reasonable time-frame.

4.2 Shape searching for de novo (-)-Englerin A Mimetics

Parts of this section are published as: Shape Similarity by Fractal Dimensionality: An Application in the de novo Design of (-)-Englerin A Mimetics. [403]
Authors: Lukas Friedrich, Ryan Byrne, Michael Mederos y Schnitzler, Aaron Treder, Inderjeet Singh, Christoph Bauer, Thomas Gudermann, Ursula Storch and Gisbert Schneider

Summary Continuing earlier work [330] with the natural product (-)-Englerin A, we sought to determine whether shape similarity could work to enrich a group of de novo designed compounds build on a pharmacophoric template. We utilised the GFD descriptor, to see whether a coarse-grained measure without local shape environments could capture useful information for the selection of novel compounds. Given the diversity of the compounds retrieved from our de novo designed library, we found that a global, shape-based approach confers similar enrichment, while capturing diverse actives than a fingerprint approach. We identified a promising hit compound representing a scaffold hop from the template natural product, and conducted a small shape-focused SAR to determine the tolerance to substitution of the identified scaffold.

4.2.1 Similarity and SAR studies

Retrospective analysis We begin with an analysis of three data sets; (i) the initial de novo design set (323 computer-generated molecules), (ii) the thirty top-ranked compounds in terms of global fractal dimensionality distance (GFD distance), and (iii) the thirty top-ranked compounds compounds according to their topological pharmacophore similarity (CATS distance) to (-)-Englerin A. Set (iii) was included to compare the GFD ranking approach with the CATS approach described previously. [330] As a first approach, we extracted the molecular scaffolds (BMS) of these compounds and analysed their scaffold diversity in terms of the pairwise Jaccard-Tanimoto coefficient (Tc) based on Morgan structural fingerprints (radius = 2; equivalent to ECFP4). The 323 initial de novo designs consisted of 152 unique scaffolds (47%) with high diversity (Tc = 0.18). The 30 top-ranked molecules according to glsgfd distance contained 24 unique (80%) and diverse (Tc = 0.17) scaffolds, whereas the 30 top-ranked compounds by CATS distance comprised 19 unique scaffolds

(63%) with slightly lower diversity (Tc = 0.24). Only two scaffolds were present in both top-ranking sets.

Second, we employed an experimentally-validated target-prediction software (SPiDER [405, 406]) to provide an estimate of the likelihood of a given compound being active against the target family 'Transient Receptor Potential Ion Channel'. The top 30 compounds retrieved by screening with the GFD, USR, SHAEP, and ECFP4 methods were analysed to determine their SPiDER predicted activity (number of compounds with an annotated $p < 0.05$) for the target family, and the proportion and diversity of the unique molecular scaffolds for the predicted active compounds was analysed. We also performed activity prediction and diversity analysis for the library in its entirety (predicted actives = 0.25, scaffold diversity = 0.47).

GFD retrieved 10 compounds predicted as active, each with a unique scaffold (predicted actives = 10, proportion of unique scaffolds = 1.0, diversity of unique scaffolds (pairwise Tc) = 0.22). The SHAEP approach retrieved fewer predicted-active compounds, also all having unique scaffolds (6, 1.0, 0.21). USR retrieved the same number of predicted actives as the SHAEP approach, with fewer unique, but highly diverse, retrieved scaffolds (6, 0.66, 0.12). ECFP4 retrieved the same number of predicted actives as GFD, but with fewer, less diverse, unique scaffolds (10, 0.8, 0.33). Given that topological approaches were used in the processes of library generation and target prediction, it is corroborative that the GFD approach, which treats sub-structural information implicitly, achieved a similar predicted-active retrieval performance under evaluation with topological methods.

In summary, SHAEP and USR have slightly poorer-than-random performance in terms of proportion of predicted actives in their top-ranked lists (0.2 for each), with variation in number and diversity of retrieved scaffolds. ECFP4 and GFD retrieve an identical number of predicted actives, with GFD having a higher, and highest-overall, number of unique molecular scaffolds in the predicted active compounds retrieved. Calculated physicochemical properties of the GFD top-ranked compounds were comparable to both the initial de novo design set (DOGS) and the pharmacophore top-ranked compounds (CATS) (see Appendix 1, Figure 1).

Prospective analysis For our prospective application, we selected the thirty top-ranked compounds according to their GFD distance, and

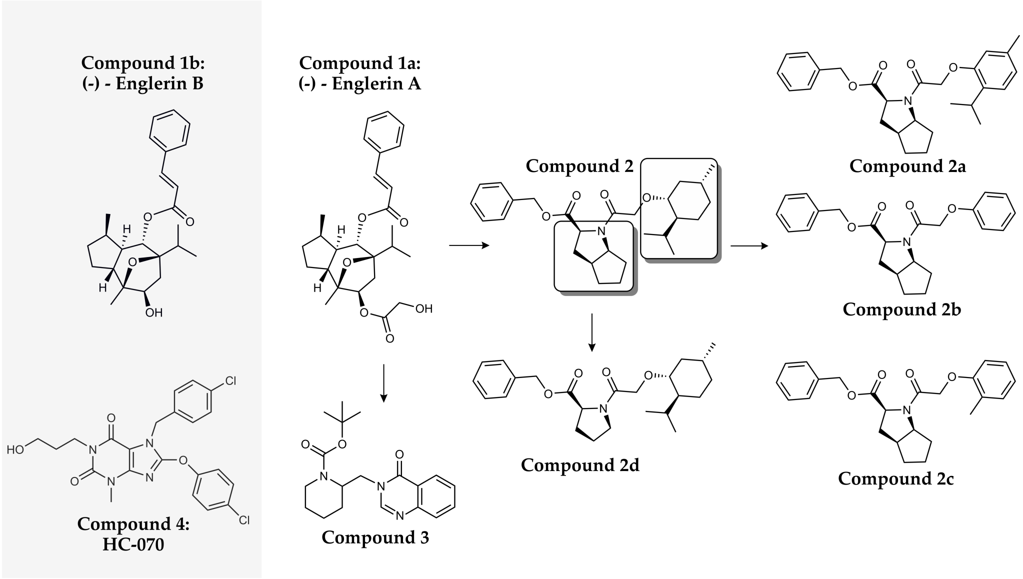


Figure 4.5: Illustration of workflow for search for synthetic analogues of compound **1a**, the natural product (-)-Englerin A (compound **1a**). Our active synthetic analogue, compound **2**, identified through GFD screening against a library of synthetically-accessible compounds as described in section 3.3, GFD Virtual Screening, and SAR, was developed into a small series based on two hypotheses. One was that the menthol group was possibly crucial for activity, but that altering it might allow for greater TRP subtype selectivity. This led to compounds **2a**, **2b**, and **2c**, testing the tolerance to an aromatic replacement, of similar shape, a more simple ring structure, and the tolerance for substituents, respectively, with thymol, phenol, and cresol substituents. The secondary hypothesis was that 'cutting' the ring structure to form a proline-backbone, would allow for a better shape match, based on visual assessment of a flexible alignment, and might therefore result in better shape similarity to compound **1a**, leading to compound **2d**. In addition, this allowed us to explore the effect of reduced backbone rigidity in this series of compounds. Compound **3** has only a weak inhibitory effect on TRPC4. Compound **4** is the most potent known inhibitor of TRPC4 and is included as a reference for the shape-based comparison. Synthesis and selection of SAR study compounds was carried out by Dr. L. Friedrich.

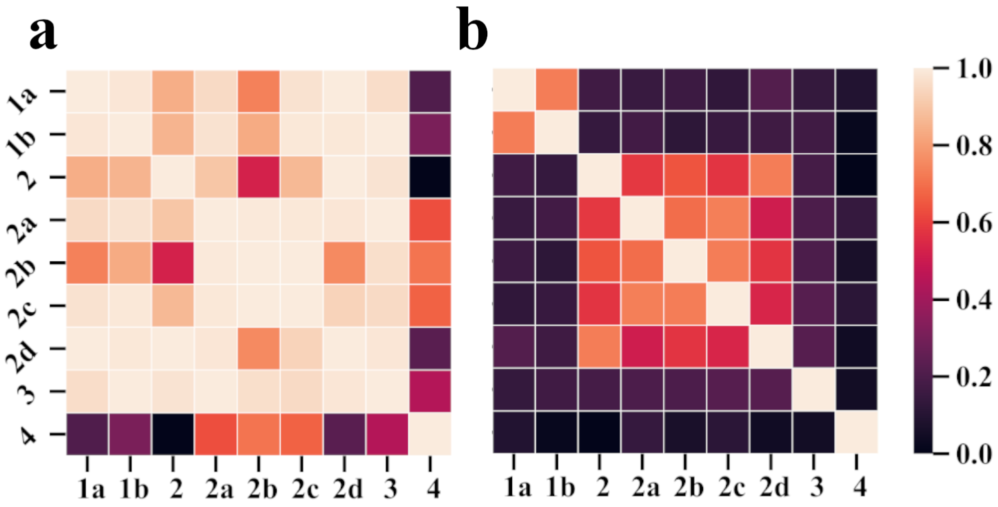


Figure 4.6: Min-max scaled pairwise similarity matrix for compounds **1**, **2**, **2a–d**, **3** and **4** under GFD and FDFP8. Graph a shows the maximum pairwise similarity (complement of Euclidean distance) under GFD, and b under FDFP8 (Dice similarity, 1024-bit representation), for a series of ten conformers of each compound. As such, the second square in the first row of graph a details the maximum similarity obtained when comparing ten conformers of compound **1a** to the ten conformers generated for **1b**, for instance. This is then scaled by the maximum and minimum similarities obtained for the entire matrix (barring the diagonal), to place the result into a useful range for visual inspection of graphs a and b, as they have different innate scales owing to their basis in different descriptor-coefficient pairings. As such, values here cannot be directly compared with those described elsewhere in this work. In general, we see that GFD has higher average pairwise similarities (0.83±0.26) than FDFP8 (0.37±0.32). Compound identifiers are as per Figure 4.5, opposite.

utilised computational target prediction to further refine our selection. Of these, nine had p values ≤ 0.05 for the target class 'TRP Ion Channel'. We selected compounds **2** and **3** for synthesis and bioactivity evaluation, considering their synthesisability and building block availability (Appendix, Section Figure 2).

Based on the results of this initial study, and on the detailed characterisation of the interactions of compound **2** with a subset of TRP ion channels carried out by our collaborators, we proceeded to generate a small structure-activity relationship series, as illustrated in Figure 4.5, to determine the importance of two structural features. Given the novelty of the menthol moiety in inhibitors of TRPC4, we sought to determine whether substitution of the menthol moiety in compound **2** is tolerated, and whether modifications would increase selectivity for TRPC4 vs. TRPM8 inhibition. Tolerance of substitution of the cyclopentapyrrole for a pyrrolidine was assessed. One suggestion, based on visual observation by an expert medicinal chemist, was that such a 'ring cut' would improve the quality of a generated overlap between compounds **1a** and **2**. In addition, this substitution would substantially alter the rigidity of the molecule. All selection and synthesis was performed by Dr. L. Friedrich.

4.2.2 Comparative similarity analysis

To compare the newly-selected SAR compounds to the original template, and the identified inhibitor, we adopted both of the FD methods discussed thus far, GFD and FDFP. For each compound in the set described in Figure 4.5, we generated a washed structure (subsection 3.1.3) and a set of ten diverse conformers using the procedure described in subsection 3.1.4, to ensure an adequate sampling of conformational space. We then constructed separate pairwise similarity matrices for each metric. We calculate GFD and FDFP8 descriptors for each conformer generated, and then perform a version of the max-fusion approach previously discussed Figure 3.4. For each pair of molecules, this results in a single number for each method, which represents the maximal similarity observed between the set of conformers for each molecule, under a given method and comparison coefficient. For GFD, this is the complement of the Euclidean distance (Equation 3.3). For FDFP8, the Dice coefficient, utilising a 1024-bit vector representation. Finally, each matrix is further processed by use of a min-max scaling procedure (Equation 3.21). As method-coefficient combinations result in different natural ranges for the resulting similarity values, this procedure is performed to allow

Table 4.13: Compound-wise average similarity for Figure 4.6. Diagonal is masked, i.e. self-similarity is removed from the calculation. We see that average similarity values are substantially higher for GFD than FDFP.

Compound ID	Mean GFD Similarity (SD)	Mean FDFP Similarity (SD)
1a	0.83(0.25)	0.22(0.19)
1b	0.87(0.22)	0.21(0.20)
2	0.75(0.31)	0.38(0.26)
2a	0.93(0.12)	0.40(0.24)
2b	0.82(0.16)	0.40(0.27)
2c	0.92(0.10)	0.40(0.25)
2d	0.86(0.25)	0.38(0.23)
3	0.91(0.18)	0.17(0.05)
4	0.40(0.24)	0.07(0.04)

for approximate visual comparison, and more meaningful numerical comparisons, between methods, to highlight the differences between approaches.

In Figure 4.6, we see a visualisation of these scaled GFD and FDFP8 pairwise similarity matrices. The GFD similarity matrix has a notably higher and less variable (0.83±0.26) average overall pairwise similarity than that observed with FDFP8 (0.37±0.32). This is also evident, graphically, in the relative colouration of the two matrices.

If we consider the overall similarity of each compound to the remainder (see Table 4.13), we find that average GFD similarity is substantially higher in each instance. In each case, compound **4** is most dissimilar from the other compounds. For GFD, compounds **1a**,**b** are essentially indistinguishable (similarity = 0.99). This effect is not replicated with FDFP (0.71). Compound **2a** is more similar to compound **4** (0.80) than its parent molecule, compound **2** (0.68), and approximately equally similar to the original template, compound **1a**) (0.82). The corresponding values for FDFP8 are 0.64, 0.03, and 0.11. Under both methods, compounds **2a**, **2b** and **2c** are more similar (GFD: 0.99, FDFP: 0.71) to each other than to **2d**, the proline derivative (0.93, 0.44). In general, FDFP seems less permissive than GFD, considering the relative similarity values to compounds **1a**,**b**, **3**, and **135** noted for each approach.

Table 4.14: Activity data for the template molecule, **1a**, and our GFD-selected compounds, **2** and **3**. n.d. indicates that a measurement was not determined.

Target	Comp 1a	Comp 2	Comp 3
TRPC4 (IC_{50})	0.011 µM	>100 µM	>100 µM
TRPC4 (EC_{50})	No effect	5.1±1.8 µM	>100 µM
TRPM8 (IC_{50})	3 µM	1.8±1.1 µM	>10 µM
TRPA1 (IC_{50})	2.62 µM	>100 µM	n.d.
TRPV3 (IC_{50})	2.84 µM	>100 µM	n.d.
TRPV4 (IC_{50})	3.91 µM	39±1µM	n.d

Table 4.15: Inhibition data for TRPM8 and TRPC4 for the series of SAR compounds developed from compound **2**. Asterisk indicates that standard deviations are not yet available for these data as of time of writing.

	TRPM8 IC_{50} (µM)	TRPC4 Inhibition (% at 10 µM)
Comp 2	1.8±1.1	49.6±9.13
Comp 2a	3.3*	30.6±7.32
Comp 2b	10*	26.4±17.1
Comp 2c	2.5*	31.9±14.6
Comp 2d	2.7*	15.3±17.8

4.2.3 Bioactivity Results

To assess their bioactivity profiles, compounds **2** and **3** were profiled in several TRP assays, in which compound **1a** showed activity (TRPC4, TRPM8, TRPA1, TRPV3, TRPV4). Since compound **1a** is a potent TRPC4 channel activator, we analysed the modulatory effects of compounds **2** and **3** on TRPC4 channels. Compound **3** had only a weak inhibitory effect of $\leq$ 20% on TRPC4 currents at a concentration of 100 µM performing electrophysiological whole-cell measurements with TRPC4 overexpressing HEK293 cells (section 5). In contrast, compound **2** showed inhibitory effects on TRPC4 channels in the same electrophysiological assay (Figure 2). Compound **1a** was used to elicit maximal TRPC4 currents. Application of stepwise increasing compound **2** concentrations in the presence of compound **1a** decreased the compound **1a**-induced TRPC4 currents. As a control, **1a** was applied for a second time inducing maximal TRPC4 currents which were used for normalization. The summary of the maximal outward currents induced by **1a** in the presence of compound **2** reveals an IC_{50} for compound **2** of 5.1±0.8 µM (K_i = 0.9 µM). Thus, we could identify compound **2** as a novel TRPC4 channel blocker, and the first known menthol-containing compound to interact significantly with that channel. This was confirmed by substructure searching in the ChEMBL and PubChem databases.

Considering the inhibitory data in Table 4.15, we see that all derivatives have a lower inhibition of TRPC4 at 10 µM than the template compound **2**. Compounds **2a** and **2c** see a mild-to-moderate loss of activity for both TRPC4 and TRPM8. Compound **2b** has further reductions in each.The same series of changes somewhat reduces TRPC4 inhibitory ability also. TRPC4 inhibition was largely abolished for the 'ring-cut' derivative, compound **2d**, whilst the reduction in TRPM8 inhibitory effect is minimal. Although this compound was closer to compound **2** under both FD approaches (GFD and FDFP8) than it was to the menthol derivatives, it sees a much more substantial reduction in TRPC4 inhibitory effect. Both GFD and FDFP8 identify compound **2d** as the least similar derivative to compound **4**, but this applies equally to compound **2** itself.

4.2.4 Discussion

Therefore, based on this small study, it seems that substitutions of the menthol group are largely tolerated for TRPM8 inhibitory activity, with the least similar compound (**2b**) from this series of derivatives (**2a**, **2b** and **2c**) under GFD to compounds **1a** and **2** showing the most substantial reduction in effect. As such, this study emphasises that both GFD and FDFP8 are useful tools for scaffold-hopping and exploring options which might not occur to a chemist, but equally that any similarity method 'without intelligence' can sometimes struggle to identify which parts of a molecule are important for bioactivity.

Combining approaches, such as integrating CATS or SPiDER can help to overcome the limitations of individual methods, but each method has its own strengths as well. This issue is, of course, not unique to our approach, but is a more general consideration when attempting to navigate biological space with a chemical map. Although we were able to identify a reasonably potent, novel scaffold, a similar study where **2** was replaced in the library with one of its derivatives would quite possibly be registered as a failure; as such, the presence of activity cliffs [33] in chemical space continues to confound attempts at rigorous method evaluation with prospective screening. Quantifying the strengths and weaknesses of developed methods is essential to deploying them appropriately when we come to ‘real-world’ tasks, but without a broad statistical basis, and good approaches to compare developed descriptors and strategies, it remains difficult to see the cliff ahead of time.

4.3 Prospective study

Summary Having conducted a focused study investigating the potential of GFD as a coarse screening metric, we decided on a larger-scale profiling of FDFP, based on improved performance in the retrospective tasks (see subsection 4.1.1), and finer discrimination between groups in the natural product study (see subsection 4.2.2). As such, we performed a large-scale virtual screening for seven targets of pharmaceutical interest, for which multiple high-potency compounds were previously known, and which we have previously adopted in our lab as a diverse set for the benchmarking of virtual screening methods.

A common criticism of LBVS validation efforts is that, owing to the dearth of established benchmarking standards, authors have a large degree of latitude in defining criteria for success. In the retrospective study, we adopted an established benchmarking platform, to limit opportunities to introduce bias. To address this in the prospective case, we opted for a strategy which, while likely reducing the number of retrieved actives, would allow us to determine the relationship between similarity and activity for our target sets over a larger range. Activity cliffs, where small shifts in chemical space can dramatically curtail bioactivity, are a known hazard in QSAR construction, and, equally, to the utility of similarity-screening methods [428–431]. Given the novelty of the FDFP approach, we attempt to quantify this 'dropping-off' in activity, by selecting compounds with a range of FDFP8-Dice similarity values to known templates (see subsection 3.4.1). This approach allows us to assess the generalisability of insights inferred from our retrospective data to a prospective, shape-driven screening approach. In so doing, we likely reduce the proportion of active compounds retrieved, but create a high-quality dataset and testing framework, allowing for further analysis of the utility of the approaches discussed thus far.

Overall, 28/130 (22%) compounds tested were active at their intended target, with this figure rising to 42% for retrieved compounds over the similarity threshold. Of these 28 compounds, 22 had BMS which were not known for their respective targets, based on a survey of publicly-available biochemical data. Using these activity and similarity data, we utilised a probability model trained on data retained from our benchmarking study, and assessed the applicability of insights learned from these to our FDFP-ordered space. We found that the FDFP8 model generalised well in this case, with others showing reduced performance, and that the method retrieves compounds which are not predicted to be active by two popular target-prediction software.

4.3.1 Target overview

We selected seven targets of pharmaceutical interest, differing substantially in their structure and function, which we have previously used in-house for LBVS method validation (unpublished) (Table 3.2).

Adenosine A2a (A2A), a GPCR with a somewhat unusual binding pocket [432] which results in extended ligand conformations, is a target of interest in several immunologic [433], and neurological disorders, including Parkinson's disease [432]. Additional reports have suggested that A2a inhibition might increase the efficacy of multiple-checkpoint inhibitors [434]. Cannabinoid-receptor 1 (CB1) is another GPCR, and a native target of endogenous cannabinoids. Interestingly, this receptor has been targeted with antagonists, inverse agonists, and agonists, for varying reasons. Inverse agonists, such as Rimonabant, were found to aid weight-loss and smoking-cessation efforts, although it was discontinued in severak markets owing to negative psychiatric effects [435]. Agonists, such as Dronabinol, have powerful anti-emetic and anti-nauseatic effects, and also act as appetite stimulants [436]. The metabotropic glutamate receptor 5 (mGluR5) is a GPCR target of considerable interest in the development of anxiolytic, antidepressant and anti-addictive medications [437]. Inhibitors of this receptor also seem to have a specific impact on suicidal ideation [438].

The glucocorticoid receptor 1 (GRG) is a cytosolic nuclear receptor, and an a target of endogenous cortisol, involved in the regulation of gene transcription throughout the body. As such, regulating their activity allows for modulation of inflammatory and immunological processes. The majority of approved drugs are agonists, such as dexamethasone, but antagonists, such as mifepristone, are also of medical interest [439–441]. The peroxisome proliferator-activated receptor delta (PPAR-δ) is a nuclear hormone receptor. The family of receptors have established use as targets of antidiabetic and antihypercholesterolaemic drugs (fibrates, thiazolidindiones), although PPAR–γ activation has been associated with adverse outcomes, and a poor side-effect profile. Although it remains controversial [442, 443] , PPAR-δ agonists have been proposed as antidiabetic, anti-neurodegenerative, and anti-atheroschlerotic agents. Additional proposed activities are as anti-inflammatory agents [444]. However, some sources [445–447] have suggested a proliferative effect for PPAR-δ agonists in colon cancer cell lines.

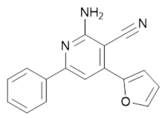
Mitogen-activated protein kinase 8, or c-jun N-terminal kinase (JNK1) plays an important, and wide-ranging role in the regulation of immune and inflammatory responses, through interactions with c_Jun, TNF-α, and NF-κB [448], and is of interest as an anti-proliferative and anti-neoplastic target of small molecule inhibitors [449]. Some reports suggest that JNK1 inhibitors could be of use as neuroprotective agents in cases of ischaemic stroke [450].

Finally, the proto-oncogene serine/threonine-protein kinase (PIM1) is involved in cytokine signalling and transcription regulation. Inhibitors of this kinase have been proposed as novel approaches to inhibiting amyloid plaque formation [451], but interest has primarily come from investigations into their effects as antineoplastic agents [452], specifically for head-and-neck cancers, myeloma, lymphoma, and myeloid leukaemia [453]. Given this, it remains the focus of active development efforts, with nearly 40 patents published between 2009-2013, for example [454–456].

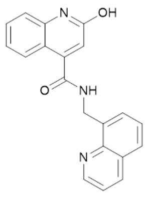
4.3.2 Compound selection

For each of the targets discussed in subsection 4.3.1, a subset of compounds with a 'standard value' $\geq$7 were extracted from the ChEMBL database (see subsection 3.4.1), and limited to those compounds displaying the desired agonistic or antagonistic effect. We sampled compounds with a range of similarity values, choosing top hits, and those around the 1000th rank. Other approaches, such as binning similarity values, were considered, but having sufficient coverage of the distributions for seven targets would require a significantly larger experiment. Our simplistic approach allows for dense coverage of the space of most interest in typical LBVS efforts, i.e. the top-ranked compounds, whilst allowing us to describe how the likelihood of activity tails off over this range. 131 compounds had their activities profiled overall; approximately 20 per target, depending on eventual availability in sufficient quantity at the supplier.

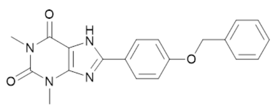
Activity data After assay as described in subsection 3.4.1, activity data were obtained for 131 compounds. One compound displayed non-specific interactions, and has been excluded from further analysis as its activity could not be determined. As discussed in subsection 3.4.1, we discretise the absolute percentage inhibition/activation values to determine whether or not a compound is active, at the 30% threshold, except for those compounds tested on PPAR-δ, where an EC_{50} criterion is used.



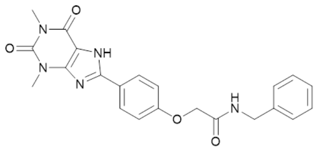
Cmpd. **5**



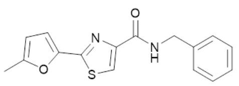
Cmpd. **6**



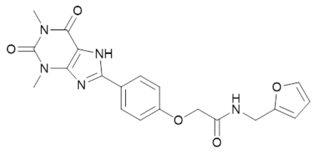
Cmpd. **7**



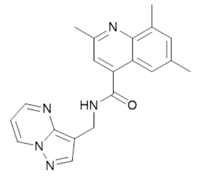
Cmpd. **8**



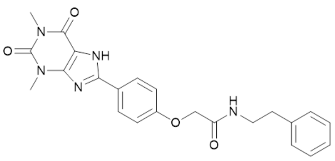
Cmpd. **9**



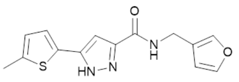
Cmpd. **10**



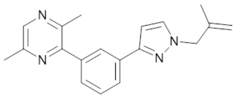
Cmpd. **11**



Cmpd. **12**

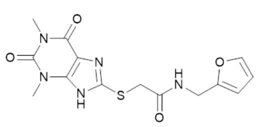


Cmpd. **13**

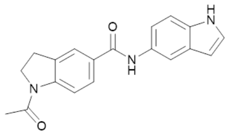


Cmpd. **14**

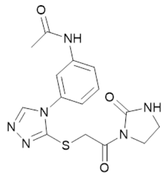
Figure 4.7: Compounds purchased for testing against the A2a receptor



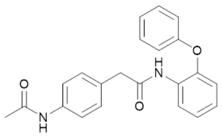
Cmpd. **15**



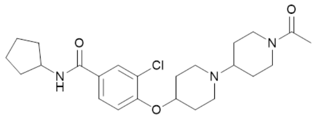
Cmpd. **16**



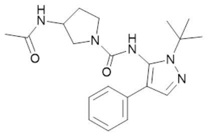
Cmpd. **17**



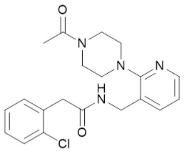
Cmpd. **18**



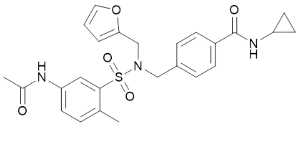
Cmpd. **19**



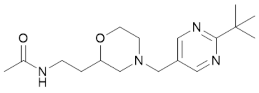
Cmpd. **20**



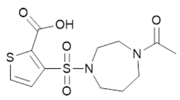
Cmpd. **21**



Cmpd. **22**

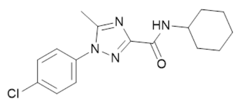


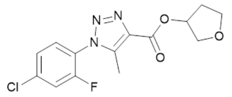
Cmpd. **23**



Cmpd. **24**

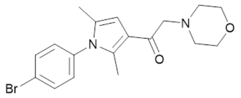
Figure 4.7: Cont.

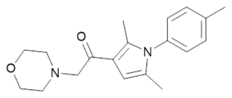




Cmpd. **25**

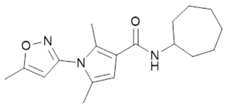
Cmpd. **26**

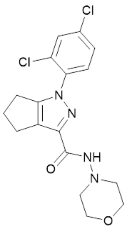




Cmpd. **27**

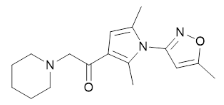
Cmpd. **28**

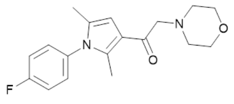




Cmpd. **29**

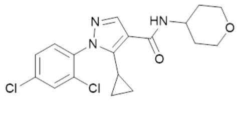
Cmpd. **30**

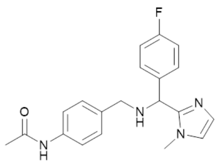




Cmpd. **31**

Cmpd. **32**

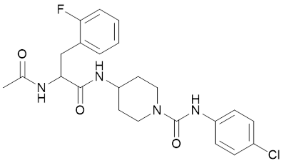




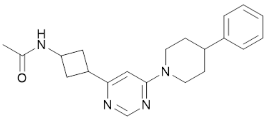
Cmpd. **33**

Cmpd. **34**

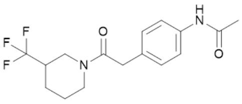
Figure 4.8: Compounds purchased for testing against the CB1 receptor



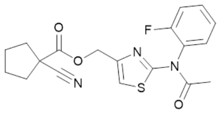
Cmpd. **35**



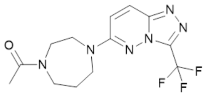
Cmpd. **36**



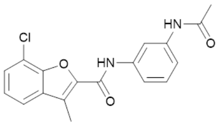
Cmpd. **37**



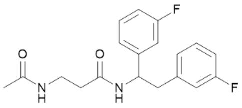
Cmpd. **38**



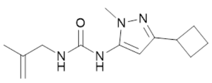
Cmpd. **39**



Cmpd. **40**

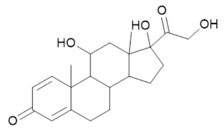


Cmpd. **41**

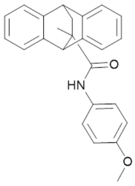


Cmpd. **42**

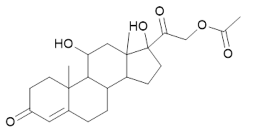
Figure 4.8: Cont.



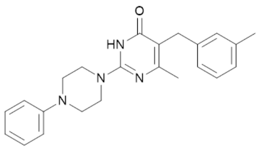
Cmpd. **43**



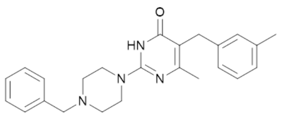
Cmpd. **44**



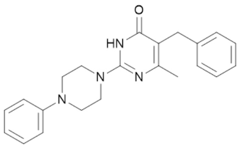
Cmpd. **45**



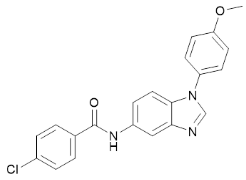
Cmpd. **46**



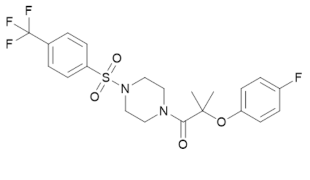
Cmpd. **47**



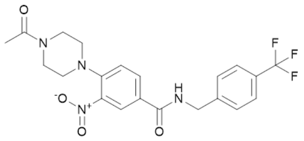
Cmpd. **48**



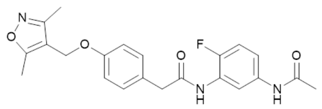
Cmpd. **49**



Cmpd. **50**

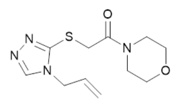


Cmpd. **51**

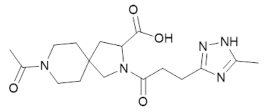


Cmpd. **52**

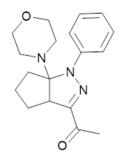
Figure 4.9: Compounds purchased for testing against the GRG receptor



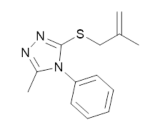
Cmpd. **53**



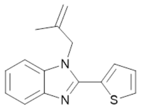
Cmpd. **54**



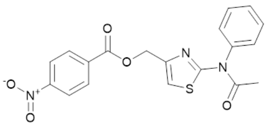
Cmpd. **55**



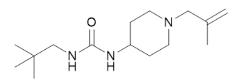
Cmpd. **56**



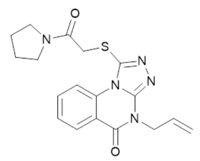
Cmpd. **57**



Cmpd. **58**

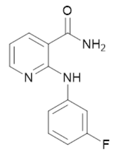


Cmpd. **59**

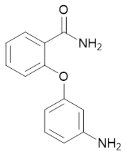


Cmpd. **60**

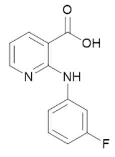
Figure 4.9: Cont.



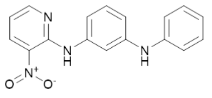
Cmpd. **61**



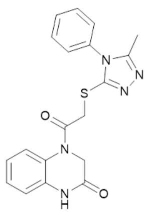
Cmpd. **62**



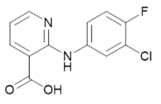
Cmpd. **63**



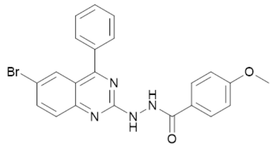
Cmpd. **64**



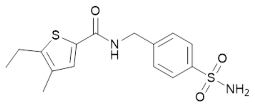
Cmpd. **65**



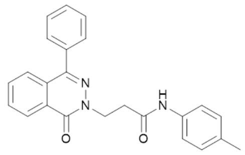
Cmpd. **66**



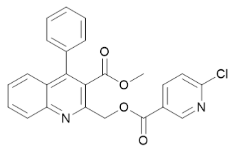
Cmpd. **67**



Cmpd. **68**

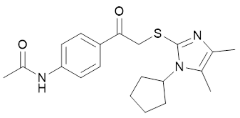


Cmpd. **69**

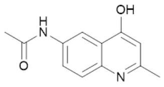


Cmpd. **70**

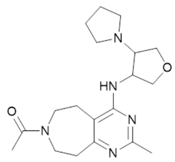
Figure 4.10: Compounds purchased for testing against the JNK1 receptor



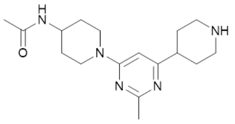
Cmpd. **71**



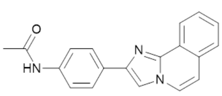
Cmpd. **72**



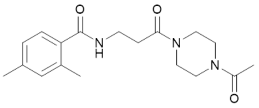
Cmpd. **73**



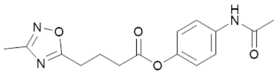
Cmpd. **74**



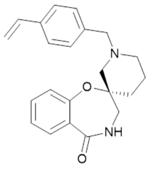
Cmpd. **75**



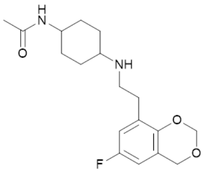
Cmpd. **76**



Cmpd. **77**

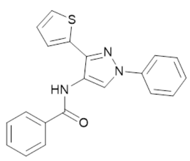


Cmpd. **78**

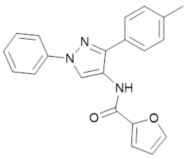


Cmpd. **79**

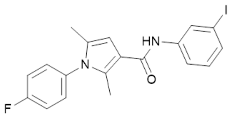
Figure 4.10: Cont.



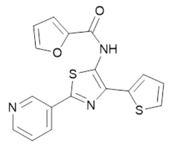
Cmpd. **80**



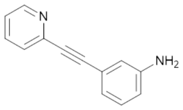
Cmpd. **81**



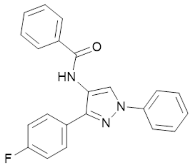
Cmpd. **82**



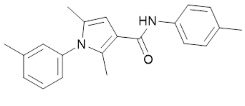
Cmpd. **83**



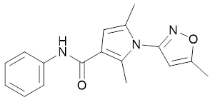
Cmpd. **84**



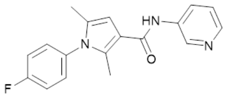
Cmpd. **85**



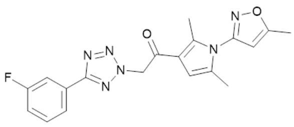
Cmpd. **86**



Cmpd. **87**

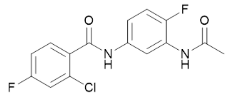


Cmpd. **88**

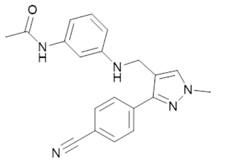


Cmpd. **89**

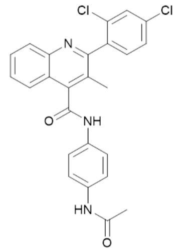
Figure 4.11: Compounds purchased for testing against the mGluR5 receptor



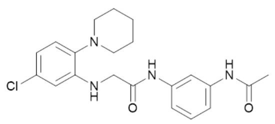
Cmpd. **90**



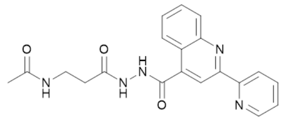
Cmpd. **91**



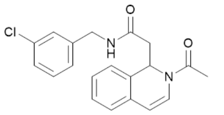
Cmpd. **92**



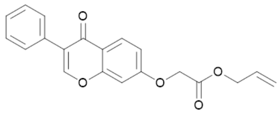
Cmpd. **93**



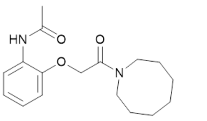
Cmpd. **94**



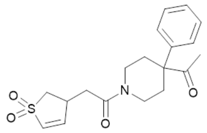
Cmpd. **95**



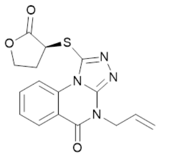
Cmpd. **96**



Cmpd. **97**

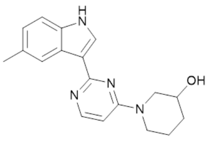


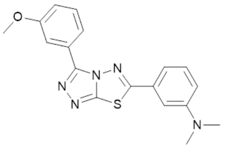
Cmpd. **98**



Cmpd. **99**

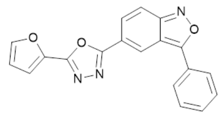
Figure 4.11: Cont.

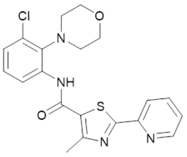




Cmpd. **100**

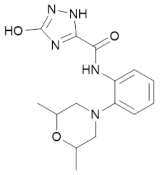
Cmpd. **101**

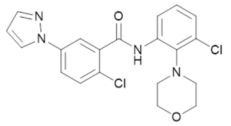




Cmpd. **102**

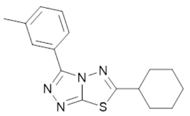
Cmpd. **103**

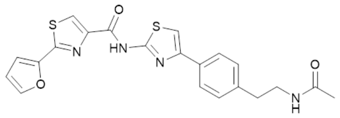




Cmpd. **104**

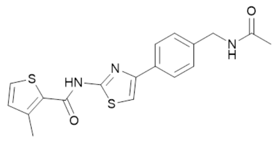
Cmpd. **105**

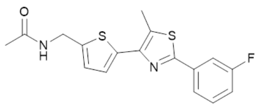




Cmpd. **106**

Cmpd. **107**

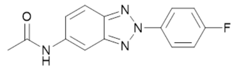




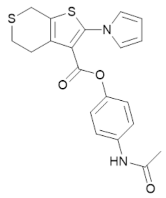
Cmpd. **108**

Cmpd. **109**

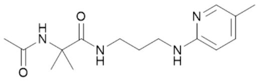
Figure 4.12: Compounds purchased for testing against the PIM1 receptor



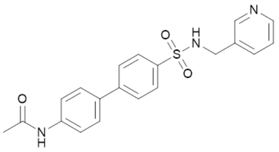
Cmpd. **110**



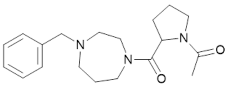
Cmpd. **111**



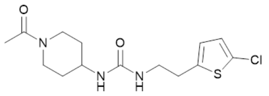
Cmpd. **112**



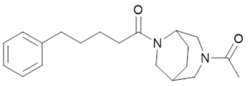
Cmpd. **113**



Cmpd. **114**

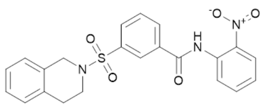


Cmpd. **115**

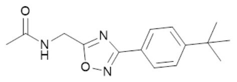


Cmpd. **116**

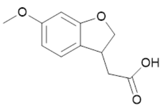
Figure 4.12: Cont.



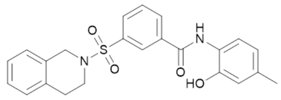
Cmpd. **117**



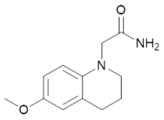
Cmpd. **118**



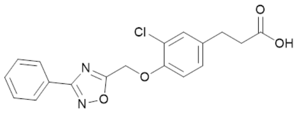
Cmpd. **119**



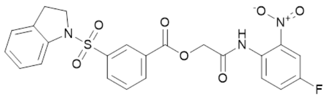
Cmpd. **120**



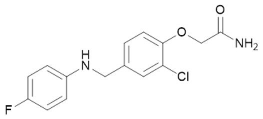
Cmpd. **121**



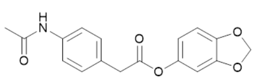
Cmpd. **122**



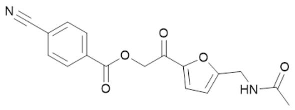
Cmpd. **123**



Cmpd. **124**

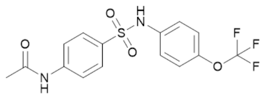


Cmpd. **125**

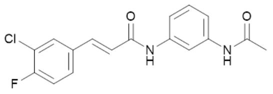


Cmpd. **126**

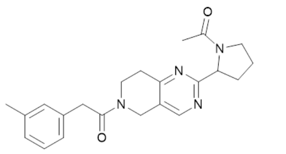
Figure 4.13: Compounds purchased for testing against the PPAR-δ receptor



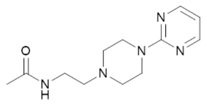
Cmpd. **127**



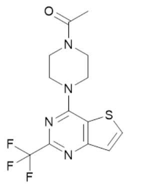
Cmpd. **128**



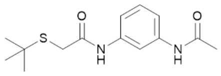
Cmpd. **129**



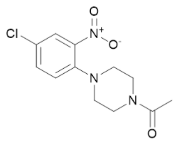
Cmpd. **130**



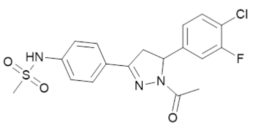
Cmpd. **131**



Cmpd. **132**



Cmpd. **133**



Cmpd. **134**

Figure 4.13: Cont.

Overall, 28 out of 130 compounds were active at their intended target, an overall hit rate of 22%. As previously discussed, we chose to sample compounds over a broad range of similarity values, based on FDFP8-Dice similarity, such that we could better define the relationship between shape-similarity and on-target activity for this novel approach. Using the cutoff threshold obtained from an isotonic model trained on the benchmarking training set (see Table 4.21), we find that, under FDFP8, we achieve a hit rate of 42% (*n*=26). This compares to 70% (*n*=12) , 66% (*n*=6) , 50% (*n*=1) and 25% (*n*=1) for ECFP6, ROCS Combination, ROCS Shape and USRCAT, respectively (See Figure 4.14). As such, the recall of our FDFP model is substantially better than was achieved for the other similarity methods considered (See Table 4.21) on this set of compounds.

A complete listing of compound activity values is found in Table 4.17 and Table 4.16. To the best of our knowledge, only two of the tested compounds were previously known as inhibitors of these targets (Compounds **5** and **43**). No significant correlation with incidence, i.e. the number of times a given compound was highly-ranked in the repeated ranking process, was observed. The number of active compounds retrieved per target varied; A2A (8), CB1 (2), GRG (4), JNK1 (2), mGluR5 (3), PIM1 (5), PPAR-δ (4). For JNK1, PIM1, and PPAR-δ (see Table 4.16, Table 4.16), subtype-selective active compounds were identified. No similar tests were carried out for the other targets profiled.

Table 4.16: Activity data, and predicted activity, collated for all compounds tested in the FDFP prospective study. Grey areas indicate that a given value is not relevant for that combination of target and compound.

Target	Mol ID	FDFP8	ECFP6	SEA	SPiDER	%	Act/Inact
A2a	**5**	0.92	1	✓	✗	-101	✓
A2a	**6**	0.88	0.74	✓	✗	-5	✗
A2a	**7**	0.86	0.78	✓	✓	-67	✓
A2a	**8**	0.86	0.72	✓	✓	-100	✓
A2a	**9**	0.83	0.51	✓	✗	-50	✓
A2a	**10**	0.83	0.7	✓	✓	-96	✓
A2a	**11**	0.81	0.35	✗	✓	0	✗
A2a	**12**	0.81	0.7	✓	✓	-92	✓
A2a	**13**	0.8	0.4	✓	✓	-38	✓
A2a	**14**	0.74	0.41	✗	✗	0	✗
A2a	**15**	0.73	0.46	✓	✗	-2	✗
A2a	**16**	0.59	0.34	✗	✗	0	✗
A2a	**17**	0.56	0.36	✗	✗	-33.23	✓
A2a	**18**	0.56	0.4	✗	✗	0	✗
A2a	**19**	0.54	0.38	✗	✗	0	✗
A2a	**20**	0.53	0.39	✗	✓	0	✗
A2a	**21**	0.52	0.44	✗	✗	-1.07	✗
A2a	**22**	0.52	0.37	✗	✗	0	✗
A2a	**23**	0.51	0.34	✗	✗	0	✗
A2a	**24**	0.49	0.32	✗	✓	-6.34	✗
CB1	**25**	0.78	0.63	✓	✓	5	✗
CB1	**26**	0.77	0.44	✓	✗	3	✗
CB1	**27**	0.76	0.45	✓	✓	13	✗
CB1	**28**	0.76	0.4	✓	✓	12	✗
CB1	**29**	0.76	0.49	✓	✗	-41	✓
CB1	**30**	0.75	0.68	✓	✓	78	✓
CB1	**31**	0.75	0.33	✗	✗	-15	✗
CB1	**32**	0.74	0.46	✓	✗	-4	✗
CB1	**33**	0.72	0.54	✗	✓	24	✗
CB1	**34**	0.61	0.4	✗	✗	0	✗
CB1	**35**	0.6	0.43	✓	✗	12.5	✗
CB1	**36**	0.56	0.41	✗	✗	-12.46	✗
CB1	**37**	0.53	0.47	✗	✗	-0.9	✗
CB1	**38**	0.52	0.33	✗	✗	0	✗

Table 4.16 continued from previous page

Target	Mol ID	FDFP8	ECFP6	SEA	SPiDER	%	Act/Inact
CB1	**39**	0.52	0.38	✗	✗	0	✗
CB1	**40**	0.52	0.4	✗	✓	10.02	✗
CB1	**41**	0.52	0.41	✗	✗	2.35	✗
CB1	**42**	0.5	0.32	✗	✗	-5.63	✗
GRG	**43**	0.84	1	✓	✓	101	✓
GRG	**44**	0.82	0.69	✗	✓	95	✓
GRG	**45**	0.8	0.86	✓	✓	95	✓
GRG	**46**	0.74	0.54	✓	✗	2	✗
GRG	**47**	0.72	0.52	✓	✗	3	✗
GRG	**48**	0.7	0.5	✓	✗	11	✗
GRG	**49**	0.68	0.38	✗	✓	19	✗
GRG	**50**	0.61	0.49	✓	✗	12	✗
GRG	**51**	0.54	0.34	✗	✗	-1.9	✗
GRG	**52**	0.51	0.35	✗	✓	-9.26	✗
GRG	**53**	0.51	0.28	✗	✗	0	✗
GRG	**54**	0.49	0.29	✗	✗	-1.54	✗
GRG	**55**	0.49	0.34	✗	✗	-37.83	✓
GRG	**56**	0.49	0.33	✗	✗	0	✗
GRG	**57**	0.48	0.32	✗	✗	-4.28	✗
GRG	**58**	0.48	0.3	✗	✗	0	✗
GRG	**59**	0.46	0.27	✗	✗	-0.36	✗
GRG	**60**	0.44	0.35	✗	✗	0	✗
JNK1	**61**	0.89	0.52	✓	✗	30	✓
JNK1	**62**	0.87	0.41	✓	✗	0	✗
JNK1	**63**	0.85	0.58	✓	✗	0	✗
JNK1	**64**	0.85	0.36	✗	✗	6	✗
JNK1	**65**	0.8	0.35	✗	✓	0	✗
JNK1	**66**	0.77	0.43	✗	✗	0	✗
JNK1	**67**	0.76	0.47	✓	✓	0	✗
JNK1	**68**	0.74	0.6	✓	✗	13	✗
JNK1	**69**	0.71	0.43	✗	✓	15	✗
JNK1	**70**	0.71	0.45	✗	✓	36	✓
JNK1	**71**	0.56	0.35	✗	✗	1.4	✗
JNK1	**72**	0.56	0.35	✗	✗	1.2	✗
JNK1	**73**	0.54	0.29	✗	✗	1.3	✗
JNK1	**74**	0.53	0.36	✗	✗	6	✗
JNK1	**75**	0.53	0.36	✗	✗	8.6	✗
JNK1	**76**	0.51	0.33	✗	✗	5	✗

Table 4.16 continued from previous page

Target	Mol ID	FDFP8	ECFP6	SEA	SPiDER	%	Act/Inact
JNK1	**77**	0.5	0.33	✗	✗	8.13	✗
JNK1	**78**	0.49	0.26	✗	✗	0	✗
JNK1	**79**	0.48	0.35	✗	✗	8.14	✗
JNK2	**61**					23	✗
JNK2	**62**					0	✗
JNK2	**63**					0	✗
JNK2	**64**					13	✗
JNK2	**65**					7	✗
JNK2	**66**					8	✗
JNK2	**67**					2	✗
JNK2	**68**					20	✗
JNK2	**69**					14	✗
JNK2	**70**					35	✓
JNK2	**71**					0	✗
JNK2	**72**					0	✗
JNK2	**73**					0	✗
JNK2	**74**					0	✗
JNK2	**75**					0	✗
JNK2	**76**					0	✗
JNK2	**77**					0	✗
JNK2	**78**					0	✗
JNK2	**79**					0	✗
mGluR5	**80**	0.92	0.53	✗	✓	-57	✓
mGluR5	**81**	0.89	0.46	✗	✓	-26	✗
mGluR5	**82**	0.88	0.5	✗	✗	-16	✗
mGluR5	**83**	0.88	0.41	✗	✓	-20	✗
mGluR5	**84**	0.87	0.6	✓	✗	-4	✗
mGluR5	**85**	0.87	0.53	✓	✓	-40	✓
mGluR5	**86**	0.83	0.39	✗	✓	1	✗
mGluR5	**87**	0.83	0.46	✗	✓	-7	✗
mGluR5	**88**	0.82	0.52	✗	✓	-3	✗
mGluR5	**89**	0.76	0.43	✗	✗	-29	✗
mGluR5	**90**	0.63	0.43	✗	✓	0	✗
mGluR5	**91**	0.59	0.38	✗	✓	0	✗
mGluR5	**92**	0.58	0.36	✗	✓	0	✗
mGluR5	**93**	0.58	0.4	✓	✓	0	✗
mGluR5	**94**	0.57	0.37	✗	✓	0	✗
mGluR5	**95**	0.55	0.37	✗	✓	0	✗

Table 4.16 continued from previous page

Target	Mol ID	FDFP8	ECFP6	SEA	SPiDER	%	Act/Inact
mGluR5	**96**	0.5	0.3	✗	✓	0	✗
mGluR5	**97**	0.5	0.4	✓	✗	0	✗
mGluR5	**98**	0.45	0.35	✗	✓	0	✗
mGluR5	**99**	0.42	0.31	✗	✓	0	✗
PIM1	**100**	0.81	0.79	✓	✗	70	✓
PIM1	**101**	0.81	0.39	✗	✓	86	✓
PIM1	**102**	0.8	0.44	✗	✓	99	✓
PIM1	**103**	0.78	0.48	✗	✗	4	✗
PIM1	**104**	0.77	0.5	✗	✗	14	✗
PIM1	**105**	0.75	0.37	✗	✗	37	✓
PIM1	**106**	0.74	0.46	✗	✗	82	✓
PIM1	**107**	0.66	0.41	✓	✗	14.4	✗
PIM1	**108**	0.64	0.34	✗	✗	3.3	✗
PIM1	**109**	0.62	0.36	✗	✗	6	✗
PIM1	**110**	0.59	0.34	✗	✗	8	✗
PIM1	**111**	0.59	0.28	✗	✗	8.6	✗
PIM1	**112**	0.49	0.34	✗	✗	16.7	✗
PIM1	**113**	0.49	0.39	✗	✗	4.3	✗
PIM1	**114**	0.49	0.34	✗	✗	6.9	✗
PIM1	**115**	0.49	0.37	✗	✗	0	✗
PIM1	**116**	0.47	0.29	✗	✗	14	✗
PIM2	**100**					21	✗
PIM2	**101**					58	✓
PIM2	**102**					96	✓
PIM2	**103**					5	✗
PIM2	**104**					2	✗
PIM2	**105**					16	✗
PIM2	**106**					73	✓
PIM2	**107**					0	✗
PIM2	**108**					0	✗
PIM2	**109**					0	✗
PIM2	**110**					0	✗
PIM2	**111**					0	✗
PIM2	**112**					0	✗
PIM2	**113**					0	✗
PIM2	**114**					0	✗
PIM2	**115**					0	✗
PIM2	**116**					0	✗

Table 4.16 continued from previous page

Target	Mol ID	FDFP8	ECFP6	SEA	SPiDER	%	Act/Inact
PPAR-δ	**117**	0.82	0.78	✓	✗		✓
PPAR-δ	**118**	0.78	0.41	✗	✗		✗
PPAR-δ	**119**	0.77	0.5	✓	✗		✗
PPAR-δ	**120**	0.76	0.86	✓	✗		✓
PPAR-δ	**121**	0.76	0.37	✗	✗		✗
PPAR-δ	**122**	0.75	0.5	✓	✗		✓
PPAR-δ	**123**	0.75	0.48	✓	✗		✓
PPAR-δ	**124**	0.75	0.44	✓	✗		✗
PPAR-δ	**125**	0.61	0.38	✗	✗		✗
PPAR-δ	**126**	0.56	0.36	✗	✗		✗
PPAR-δ	**127**	0.56	0.41	✗	✗		✗
PPAR-δ	**128**	0.55	0.29	✗	✗		✗
PPAR-δ	**129**	0.55	0.34	✗	✗		✗
PPAR-δ	**130**	0.52	0.28	✗	✗		✗
PPAR-δ	**131**	0.51	0.32	✗	✗		✗
PPAR-δ	**132**	0.51	0.31	✗	✗		✗
PPAR-δ	**133**	0.51	0.28	✗	✗		✗
PPAR-δ	**134**	0.48	0.31	✗	✗		✗

Table 4.17: Activity of the higher-similarity PPAR-δ compounds. Where entries are not 'inactive' or 'toxic' Inactive implies that no EC$_{50}$ could be measured at 30 µM compound concentration. Compound **124** does not have a measurable EC$_{50}$, as it exhibits toxicity at 10 µM .

	PPAR-α	**PPAR-γ**	**PPAR-δ**
	EC$_{50}$ (µM)/	**EC$_{50}$ (µM)/**	**EC$_{50}$ (µM)/**
	fold-activation	**fold-activation**	**fold-activation**
119	Inactive	Inactive	Inactive
121	Inactive	Inactive	Inactive
122	5.0±0.1 / 8.6±0.1	3.3±0.3 / 6.2±0.2	12.1±0.8 / 21±1
123	Inactive	10±2 / 21±2	Inactive
120	Inactive	Inactive	0.17±0.04 / 3.9±0.1
117	Inactive	Inactive	1.9±0.4 / 2.2±0.1
124	Toxic	Toxic	Toxic
118	Inactive	Inactive	Inactive

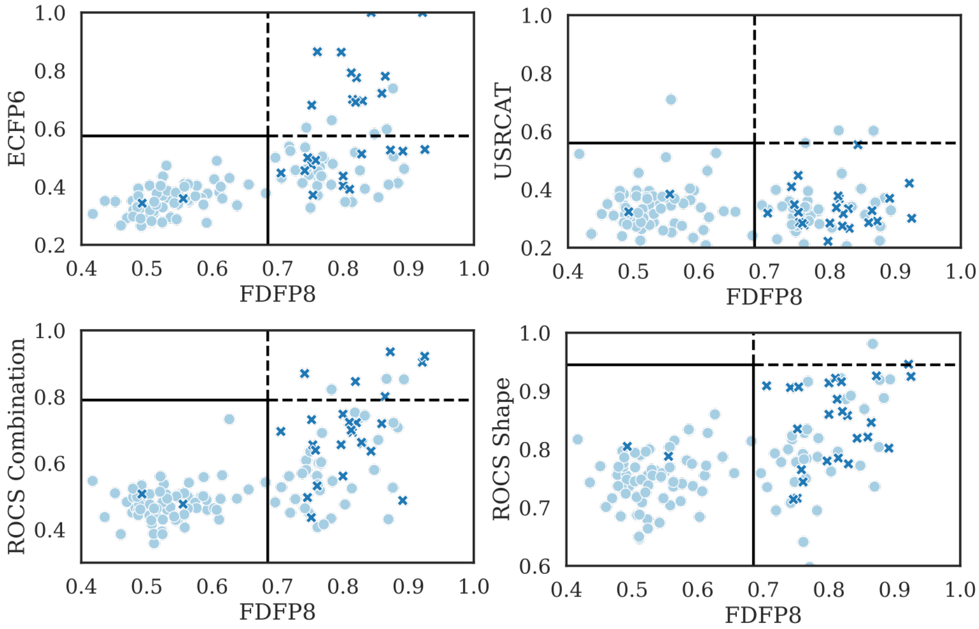


Figure 4.14: Joint distributions of maximum similarity to any template molecule for 130 compounds, under ECFP6, USRCAT, ROCS combination, and ROCS Shape, against FDFP8. Crosses mark active compounds, circles inactive. Solid lines indicate $P(active) = 0.5$ from isotonic regression models trained on the maximum similarity histograms obtained from the benchmarking study (section 3.2.2) for each similarity method. Top-left quadrant: compounds identified by y-axis similarity method but not FDFP8, top-right: compounds identified by both methods, bottom-right: compounds identified only by FDFP8, bottom-left: compounds missed by both methods. Further information on the model-training process is given in Table 4.21.

As can be seen in Table 4.18, there is considerable orthogonality between the methods profiled in terms of maximum similarity to any compound in the template set, which is also used as the representative value for the probabilistic study. We see most overlap with ROCS combination and shape (0.66). Although correlation coefficients are reasonably high and statistically-supported for FDFP8-ECFP6 (0.56) and FDFP8-ROCS combination (0.45) pairs, there is still substantial unexplained variance between methods. In addition, no method correlates significantly with USRCAT.

One important aspect of this screening is the consideration of retrieved

Table 4.18: Kendall-Tau correlation coefficients (and associated *p*-value indicators) for the pairwise comparison of maximum template similarity values for compounds profiled in the prospective study. '****' indicates a given correlation has a *p*-value ≤0.0001, 'ns' that it is not significant. N.B. the maximally-similar template may differ between methods, as shown in Figure 4.15

	ECFP6	ROCS Comb	ROCS Shape	USRCAT
FDFP8	0.56 ****	0.45 ****	0.38 ****	0.02 ns
ECFP6		0.37 ****	0.27 ****	-0.02 ns
ROCS Comb			0.66 ****	0.05 ns
ROCS Shape				0.1 ns

compound diversity. In terms of BMS, 25 novel scaffolds were identified in total, corresponding to 16 GMS (Table 4.19), when comparing the compounds studied with the highly-active compounds utilised as templates. We find that for each target, the majority of scaffolds retrieved, at both the BMS and GMS levels, are not present in the template set, indicating that this approach has discovered novel chemistry for each target. On average, the hit-rate for known scaffolds is higher, supporting the concept of privileged chemical space. We repeated the ChEMBL-scraping procedure described in subsection 3.4.1, setting the threshold of allowed activity to be equivalent to single-digit mM, which includes fragment-like binding, and determined whether our compounds shared scaffolds with these lower-activity hits (Table 4.20). We found that including this wider range of biochemical data did not significantly alter the results of our analysis, i.e., that these scaffolds were not included in this repository of publicly-available screening data, and are, for our purposes, unique, decreasing the number of previously unknown BMS with at least one active compound from 25 to 22. For GMS, we see a decrease from 16 to 13. As such, the retrieved compounds represent a diverse, novel, hit group.

Table 4.19: Annotated active compounds classified by their BMS and GMS identities, and grouped by their novelty. Here, 'unique' means that a given scaffold is not present in the collection of template molecules, with 'known' the converse. Results are given as number of scaffolds with at least one active compound/cardinality of the set of that scaffold grouping, so, for example, '7/19' indicates that seven scaffolds with at least one active compound each were identified, out of 19 scaffolds tested. For each target, a diverse pool of compounds are retrieved, with high proportions of novel BMS. The corresponding bioactivity varies widely with the distribution of similarity values for the retrieved compounds. We see that in many cases, known BMS and GMS with associated active compounds lead to inactive screening results, illustrating the difficulty of determining LBVS success.

	Unique BMS	**Known BMS**	**Unique GMS**	**Known GMS**
A2a	7/19	1/1	6/16	2/3
CB1	2/15	0/1	2/12	0/3
GRG	4/17	0/0	2/15	1/1
JNK1	2/17	0/0	2/13	0/2
mGluR5	3/18	0/1	2/14	1/4
PIM1	5/17	0/0	1/10	3/6
PPAR-δ	2/16	1/1	1/14	2/2

Table 4.20: Repetition of analysis in Table 4.19, broadening the activity threshold to include compounds with mM annotated activities against the targets assessed, a reasonable range for fragment-like hits. As we can see, the FDFP approach recovered some scaffolds which were present in the lower activity set, but was not otherwise substantially affected, indicating that the active scaffolds retrieved were truly scaffold-hops. Corresponding molecule counts: A2a, 5926; CB1, 7357; GRG, 3200; JNK1, 2403; mGluR5, 2757; PIM1, 4880; PPAR-δ, 1766.

	Unique BMS	**Known BMS**	**Unique GMS**	**Known GMS**
A2a	7/18	1/2	5/14	3/5
CB1	2/15	0/1	2/11	0/4
GRG	2/15	2/2	1/10	2/6
JNK1	1/16	1/1	1/11	1/4
mGluR5	3/17	0/2	2/12	1/6
PIM1	5/16	0/1	1/8	3/8
PPAR-δ	2/14	1/3	1/10	2/6

Illustrative examples Considering the compounds retrieved in a different fashion, we here discuss a few illustrative examples of the ‘real-world’ implications of different definitions of similarity. As our sampling procedure follows the process of repetition previously described, we have no single ‘template’ for each tested molecule. For each of the compounds, we calculate its similarity to all active compounds for its given target (as described in Table 3.2), and identify the most similar structure. We perform this procedure with FDFP8-Dice, ECFP6-Dice and ROCS-Tc combinations. This approach was taken as we are certain within reason that all of the active set molecules are indeed active, and it represents the sum of public knowledge for these targets. These compounds were not necessarily sampled as templates in the initial screening, as we chose a random sampling approach to the query set, to encourage diversity, and limit the impact of over-represented structures. To illustrate various aspects to consider with molecular similarity, we do not select the most active compounds for each target, but rather those where the disparities between each approach are most revealing.

For A2A compound **10** (96% inhibition at 10 µM) the terminal five-membered ring is not preserved in the ROCS and ECFP picks, in favour of a more rod-like substituent. The sometimes counter-intuitive nature of shape-similarity is reinforced with the neighbours of CB1 compound **29** (41% inhibition at 10 µM). Of these, the ECFP6 and FDFP8 picks seem closer to the chosen compound, in terms of complexity, and in preservation of the central five-membered ring in the latter case. However, the ECFP6-Dice similarity is approximately 0.47, which is rather low (see Figure 4.4). In contrast the ROCS-Tc pick, with a relatively inflexible central scaffold, is amongst the highest-scoring matches for the identified actives (0.64). FDFP8-Dice gives a borderline score (0.76).

The next example, compound **55** (38% inhibition at 10 µM) was chosen as it is the retrieved active with the lowest shape similarity under FDFP8-Dice (but not under ROCS-Tc) to the template structures. The compound seems reasonable, *post hoc*, given that each method contain a decalin-like scaffold with an aromatic substituent, which seems to mimic a similar scaffold in the active molecule (morpholine-substituted cyclopentapyrazole). However, the additional decoration results in a much reduced score in this case (FDFP8: 0.49, ROCS: 0.50, ECFP6: 0.34).

Finally, the differences in macro-level similarity between the templates retrieved for compounds PIM1 **100** (70% activation at 10 µM) and PPAR-δ **122** (12.1±0.8µM EC_{50}) illustrate that relatively small changes with flexi-

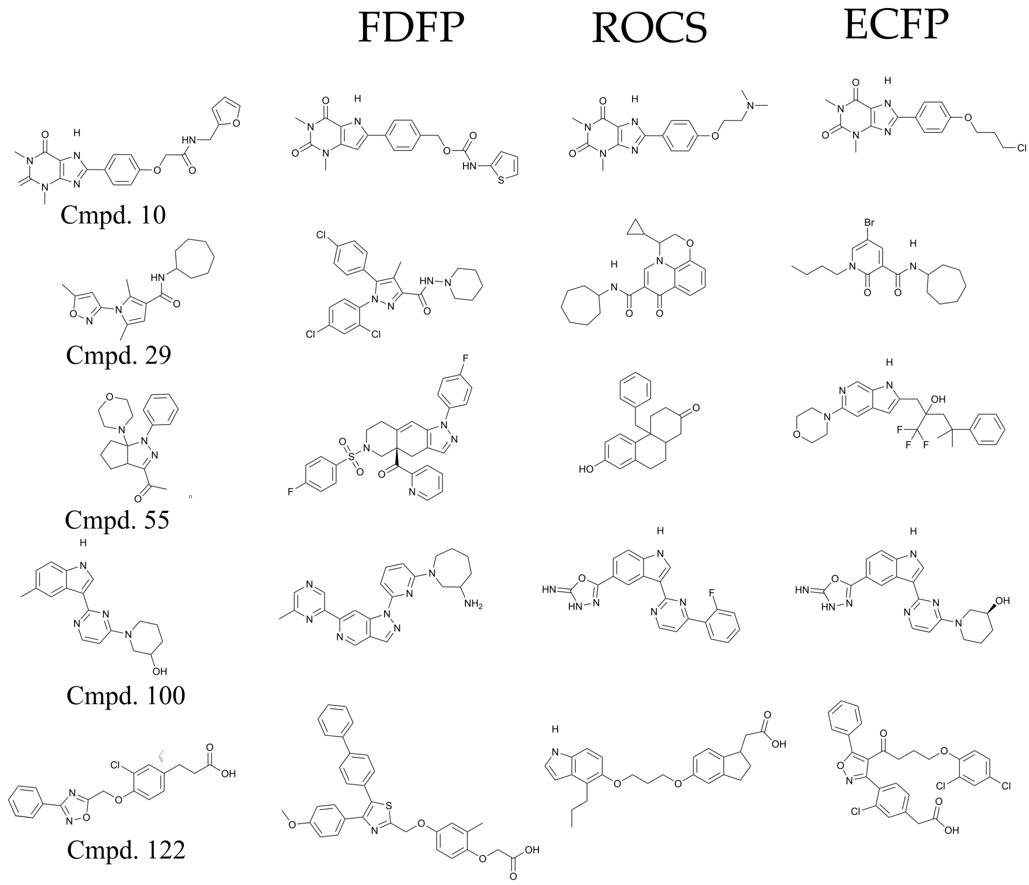


Figure 4.15: A comparison of the most similar compounds, under FDFP8, ROCS Combination, and ECFP6, to active compounds identified for targets A2a (compound **10**), CB1 (compound **29**), GRG (compound **55**), PIM1 (compound **100**) and PPAR-δ (compound **122**). Comparisons chosen to illustrate certain aspects of shape-based similarity comparison, rather than most active or novel compounds identified for each target. JNK-1 and mGluR5 are left out for the sake of brevity - In both cases, very similar molecules are returned by every method. ROCS shape, USRCAT, USR, and GFD are not discussed here, owing to their relative performance in the benchmarking experiment.

ble molecules can lead to drastically different compounds being retrieved, with both 2D and 3D-methods. In the former case, the compounds are relatively uniform, with correspondingly high similarity values (FDFP8: 0.79, ROCS: 0.70, ECFP6: 0.81). The high similarity under a topological approach is somewhat surprising, given the substantial superficial rearrangement of the molecule. In the latter, we have a more complicated picture, in which both ROCS (0.5) and ECFP6 (0.5) provide radically different candidates, neither of which scores particularly well. The ECFP6 pick contains approximately the same elements as the active compound, albeit rearranged. The ROCS compound captures the linear character of the molecule, but otherwise diverges. The FDFP8 pick contains approximately the same substructure as the chosen template, varying in heteroatom identities.

Discussion Overall, we find that the FDFP screening approach allowed for the identification of novel compounds for each of the targets profiled. Although hit-rates vary substantially, for each target at least one scaffold was identified which was previously publicly-unknown. Overall, we find that although shape similarity, as encoded here through the FDFP approach, is not always as natural a fit for chemical intuition as subgroup analysis and other 1- and 2D approaches, it allows us to retrieve novel, bioactive compounds for each target assessed, and to scan databases of tens of millions of molecules in under a minute, facilitating new discoveries, and allowing for an efficient sampling of chemical space.

While a comprehensive discussion of the chemistry of the retrieved compounds is not the subject of this work, a closer examination of the compounds retrieved, and some subjective differences between the compounds retrieved under our method and comparator topological and shape approaches, is worthwhile, as every similarity determining method builds certain biases into its model of the world. As there is no intrinsic measure of shape to act as a standard, comparison is, by necessity, somewhat subjective. Above, we highlight some surprising results with all three methods. As per Figure 4.14, we see that there are active molecules which would be 'missed' were one to use one of the other methods profiled, and that, a few consistent high-scorers aside, there are substantial differences between methods' rankings of the chemical space. Of course, we are blind to active molecules which might have been retrieved under other methods than our own, which is a limitation of prospective studies of this kind, but it serves to emphasise that our approach is somewhat orthogonal to both ROCS combination and ECFP6.

Ascertaining this based on retrospective study datasets, such as that summarised in Figure 4.2, can be difficult. Primarily this is owing to the necessary biases utilised in their construction, which renders deconvolution, and analysis, impractical. In practice, each approach has benefits; retrospective studies have much more data, prospective studies that the answer is not known ahead of time and that the dataset directly reflects the method adopted. Prospective studies also rely on biased databases; our ChEMBL sets contain extensive examples of medicinal chemical optimisation of scaffolds. Although we did not directly control for this (for example, by clustering our sets by scaffold), the FDFP approach retrieved a diverse set of active scaffolds, some of which are receptor subtype-selective. It is unclear whether this diversity is an innate benefit of shape-based methods as has been suggested [48, 211, 457, 458], or whether it varies substantially depending on the experimental setup. Our adoption of the multiple-sampling max-fusion strategy utilised in the retrospective screen may have an effect on this, and would be an interesting direction for future research.

4.3.3 Probability Models

Results Returning to the similar property principle [229, 459] which underlies most interest in virtual screening methods, we thought it appropriate to attempt quantify the extent to which we can infer similar activity under the similarity approaches considered (see section 3.2.6). We trained two regression-type machine-learning models on the similarity and activity data, generating training set 5-fold cross-validation metrics, and testing the models on the withheld activity data gained in the prospective study. The two models chosen were isotonic and logistic regression, both of which fit pseudo-sigmoidal curves, although the former makes locally-linear approximations. Both are used in probability calibration, in which the pseudo-probability output of a machine-learning model is calibrated so as to more closely reflect real-world performance.

Overall, we find that all isotonic regression models have better-than-random performance on the test set (see Table 4.21), with the ECFP, FDFP, and ROCS combination approaches demonstrating the most significant discriminatory power. In terms of recall, which reflects the proportion of actives retrieved, the ECFP and FDFP methods do significantly better than other approaches. There is relatively little shift in the threshold values observed between the isotonic and logistic approaches, however

Table 4.21: MCC, Recall and threshold values for isotonic and logistic regression models trained on distributions of similarity values (see Figure 4.4) for active and decoy sets obtained from a publicly-available benchmarking set [214], and tested on withheld data obtained from our prospective study. Threshold indicates the similarity value at which the model assigns a pseudo-probability of activity, P, ≥0.5.

Model	Similarity method	**Train**		**Test**		**Threshold**
		MCC	**Recall**	**MCC**	**Recall**	
Isotonic	**ECFP4**	**0.35±0.01**	**0.17±0.02**	0.46	0.49	0.64
	ECFP6	**0.35±0.02**	0.16±0.01	0.43	0.48	0.58
	FDFP8	0.28±0.01	**0.17±0.01**	**0.48**	**0.93**	0.69
	ROCS Combination	0.23±0.02	0.08±0.01	0.30	0.21	0.79
	ROCS Shape	0.17±0.01	0.04±0.00	0.09	0.04	0.95
	USRCAT	0.11±0.01	0.02±0.00	0.02	0.04	0.56
	USR	0.04±0.00	0.00±0.00	0.04	0.09	0.98
Logistic	**ECFP4**	**0.36±0.02**	**0.19±0.01**	**0.54**	0.43	0.71
	ECFP6	**0.36±0.02**	**0.19±0.01**	0.46	0.43	0.55
	FDFP8	0.30±0.02	0.13±0.02	0.47	**0.93**	0.68
	ROCS Combination	0.23±0.02	0.08±0.01	0.30	0.21	0.78
	ROCS Shape	0.00±0.00	0.00±0.00	0.00	0.00	1.00
	USRCAT	0.11±0.01	0.02±0.00	-0.08	0.00	0.60
	USR	0.00±0.00	0.00±0.00	0.00	0.00	1.00

the latter results in several cases where the results are suboptimal (USR, ROCS Shape), whilst only improving FDFP8 and ECFP performance by a small margin. Therefore, we chose to adopt the thresholds determined by the isotonic regression approach for further study. To compare our approach with established target prediction methods, we also utilised two publicly available tools, SEA, and SPiDER, to compare with gold-standard target-prediction performance, and determine whether such methods could capture actives with diverse shapes. Neither approach resulted in a test-set MCC greater than those observed with target-agnostic similarity values described in Table 4.21. SEA achieved an overall MCC of 0.38, with a recall of 0.64. For SPiDER, a MCC of 0.21, and recall equal to 0.5, were observed. To assess the relationship between similarity under a 2D method, such a those used to train both SEA and SPiDER, and prediction accuracy, we plotted maximal similarity for each profiled compound under FDFP8 and ECFP6 to known actives, overlaying the predictions made by the dedicated software (see Figure 4.16, and Table 4.16 for per-compound predictions).

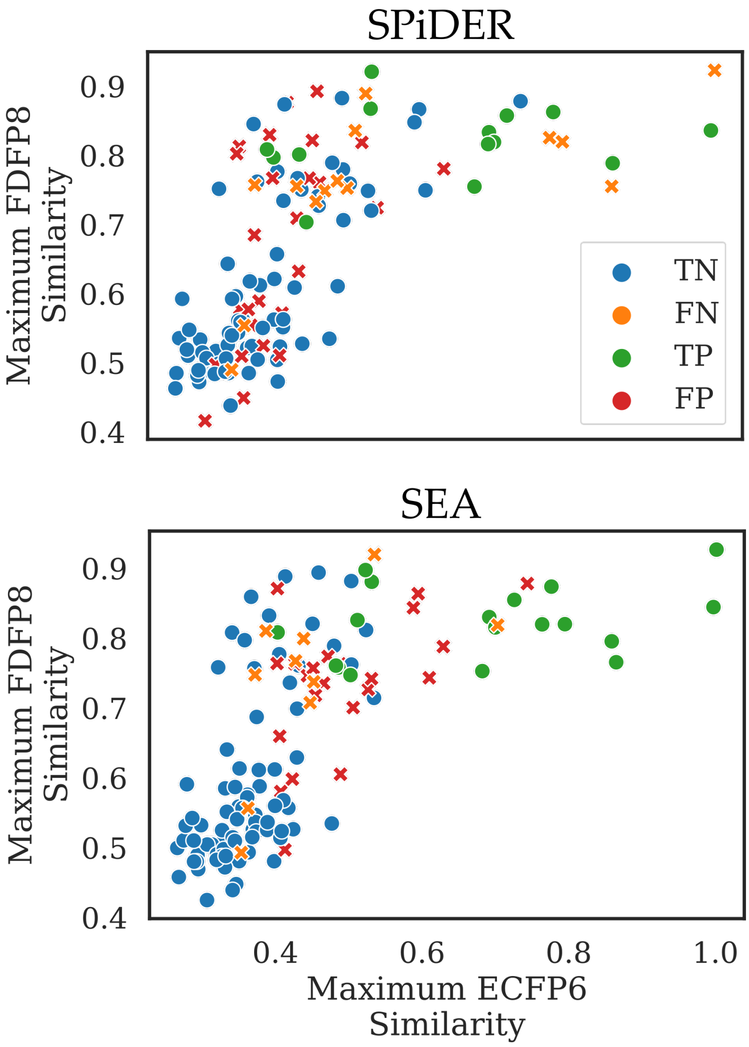


Figure 4.16: Plot of the terms of the confusion matrix for target predictions generated with the online resources SEA and SPiDER for the compounds assessed in the prospective study. Neither approach handles compounds with high shape-similarity and low ECFP6 similarity well, with many false-positives and false-negatives in that region. Interestingly, SPiDER predicts more compounds with low similarity under both approaches as active, perhaps reflecting that, in contrast to SEA, it is not built upon ECFP-type fingerprints. Both models make confident, accurate predictions for high ECFP6 similarity compounds. SEA: MCC 0.38, Recall 0.64. SPiDER: MCC 0.21, Recall 0.5. FDFP8: MCC 0.48, Recall 0.93

In general, we see that the bottom-left of each graph (low similarity under both methods), is occupied primarily by true-negative predictions in both cases, plus two false negatives. SPiDER has a significant number of false-positives in this region. The upper-left quadrant (high FDFP similarity, low ECFP similarity) sees both SPiDER and SEA perform relatively poorly compared to the top-right quadrant (high similarity under both approaches).

Discussion Both SPiDER and SEA perform significantly better where there is reasonable ECFP similarity to known templates. SPiDER seems somewhat less conservative in this regard, making false positive predictions at low ECFP4 similarity, although we did identify a few actives at that similarity level. As the data underpinning each model varies, and as SEA is a constantly-updating resource incorporating publicly available bioactivity data, it is hard to construct a secondary test-set, not incorporating shape-based screening, to establish a fair baseline performance level for each. Although SEA is trained on ECFP4 data, it achieved a lower MCC, and higher recall, than the target-agnostic approach described here. In both cases, however, the models struggled with compounds which have a high maximum FDFP and low maximum ECFP6 similarity to a known template (see Figure 4.16), demonstrating the importance of appropriate descriptor choice in the design of target prediction approaches, and suggesting that FDFP might be useful for such methods in future.

Improved enrichment with shape-based approaches in target prediction has previously been noted in the literature [106]. These ‘missed opportunities’ for traditional approaches offer considerable possibilities for the identification of novel compounds, and are a primary justification for the inclusion of shape-based approaches into virtual-screening pipelines. As we do not have ready access to the data on which SPiDER and SEA were trained, we cannot assess how our compounds are embedded within the training sets for each approach, which is a limiting factor in comparing their performances. Given the known predisposition of screening libraries towards simple, ‘linear’ [82] molecules, it seems reasonable to assume that this bias is reflected in the training data for each approach, but this should also be reflected in the libraries from which we selected compounds. An additional limitation of our model is that we do not explicitly account for the relative proportion of actives and decoys utilised in our training set. To the best of our knowledge, no study has directly considered the overall distribution of library shape characteristics when describing the performance of shape-based methods, but this could be a promising direction for future research in shape-based approaches.

One issue with our approach is that it does not take into account template molecule characteristics. Molecules with an 'unusual' shape, and those which are especially large or flexible, may well have a less clear separation of P and Q, which we have not considered here. We aim to demonstrate that empirical threshold similarity values, such as the often-quoted Tanimoto coefficient similarity of ≥ 0.7 for ECFP4, are supported by the analysis of large, diverse datasets, and can be obtained in a rigorous fashion. For ROCS shape, a similarity threshold of 0.75-0.85 has previously been discussed in the literature [48, 460], with the higher of the two employed for screening, which might suggest that our model is somewhat conservative. However, given the distribution of similarity values provided [48], this is likely due to intent rather than a more fundamental issue. We were unable to find authoritative sources for cut-off similarity values for the other methods discussed, highlighting the importance of defining a generalised approach. As a secondary matter, such approaches could be used to weight similarity values obtained from multiple sources, and give a measure of confidence to virtual screening campaigns.

None of the approaches assessed has a perfect predictive power, reflecting the complex relationship between similarity in chemical and biological spaces. Our definition of activity, at the 30% inhibition or activation threshold, may be unduly restrictive, as might testing at a single concentration, given the observed tendency for superficially very similar molecules to have multiple orders-of-magnitude variation in their IC_{50}, for instance. As such, exploring a wider range of concentration values, where experimentally feasible, could give us a better idea of the 'near-misses', compounds which have 'fallen off the activity cliff'. Regardless, our simple probabilistic models achieved gold-standard performance, and better, on this small but diverse test set, and seem promising for further exploration. Most importantly, they give us some measure of confidence in our definition of sufficient similarity, which is critical to interpreting the results of any virtual screen, as the closest compounds in a small, or insufficiently diverse, virtual screening library, may still be too far away to be confident in our ability to generalise.

4.4 Shape Screening for PIM1 Inhibitors

Summary We conducted an in-depth profile for one of the targets for which we had identified active compounds, PIM1. Although we had not made explicit attempts to promote specificity in our screening campaign, we found that the most active hit (compound **102**) in the initial profiling is amongst the most selective potent inhibitors of PIM1 for which we could find sufficient data, and substantially different from molecules disclosed to date.

Compound **102** is a potent inhibitor of PIM1, GSK-3β and DRAK-1 kinases. Comparing compound **102** to known kinase inhibitors with available selectivity data, we found that it is more selective than four out of five kinase inhibitors in the set which have been trialled in man, and amongst the upper-third of inhibitors overall. For PIM1, we found that it is amongst the most selective high-potency inhibitors known. We investigated the role of shape similiarity in determining specificity in this case.

A crystal structure for the inhibitor in complex with PIM1 was obtained, and the results analysed. Analysis revealed that compound **102** binds in the same site, and in a similar orientation, to known inhibitors of PIM1. The relative stability of FDFP8 with conformal variability, as compared to the two ROCS methods, was confirmed. A subsequent analysis of the similarity relationship between ligand and crystal shape shows a mild positive correlation.

4.4.1 Activity data

Of the seven compounds chosen for IC_{50} determination, four are low micromolar inhibitors of the PIM1 kinase, and three have sub-50 µM IC_{50} values for the Pim-2 kinase (Table 4.22). Compounds **102** and **100** are at least 5-fold selective against PIM-2. These kinases are similar at both the sequence (55% sequence identity, 67% sequence similarity) and fold levels [461].

We searched ChEMBL 25, and PubChem (July 2019) using their online portals, and were unable to find previous evidence of PIM1 inhibition for this compound. Additionally, a similarity search for its BMS on ChEMBL at 0.7 Tc with ECFP4, and a substructure search on the PubChem site, did not reveal further relevant records for PIM1. Additionally, as best we are able to ascertain, the most similar patent compounds' 'claimed' chemical space does not cover the compound [462].

Table 4.22: Determined IC_{50} values for PIM1 and PIM-2 kinases, with a selection of the compounds chosen for that target

Compound Number	Pim-2 IC_{50} (nM)	PIM1 IC_{50} (nM)
100	>50,000	5444
101	7922	1625
102	3408	484
103	>50,000	>50,000
104	>50,000	>50,000
105	>50,000	>50,000
106	6527	3046

While inhibition of both PIM1 and Pim-2 kinases is typically regarded as beneficial, with each promoting improved regulation of apoptosis, and with PIM1$^{-/-}$ Pim-2$^{-/-}$ Pim-3$^{-/-}$ knockout mice having a survivable phenotype [463], kinase-specific inhibitor scaffolds are comparatively uncommon, and have been extensively discussed as a means of reducing off-target and side-effects associated with kinase inhibitors in clinical trials [409, 464, 465].

4.4.2 Selectivity analysis

Results To better characterise the selectivity of compound **102**, a kinase panel profile was ordered from Cerep Eurofins. The compound was profiled at a concentration of 10 µM, against a panel of 58 kinases. The compound was discovered to be a potent (here defined as inhibition $\geq$70%) inhibitor of human DRAK-1 (87%) and GSK-3β (82%) kinases, in addition to PIM1. At the moderate potency level (here defined as 50%$\geq$ inhibition <70%), the compound is an inhibitor of CHK-1 (56%), CK-1γ1 (62%), and Rsk1 (56%) kinases. A further nine kinases were inhibited at a lower level (30%$\geq$ inhibition <50%) (a full list of per-kinase inhibition values is provided in Appendix 5, section 5. See Figure 4.17 for a schematic overview. A reference target-labelled map can be found in Appendix 5, Figure 3). The kinases for which compound **102** is a moderate or potent inhibitor belong to the calmodulin-dependent protein kinase (CAMK) and GMGC groups, which cover a wide range of functional space.

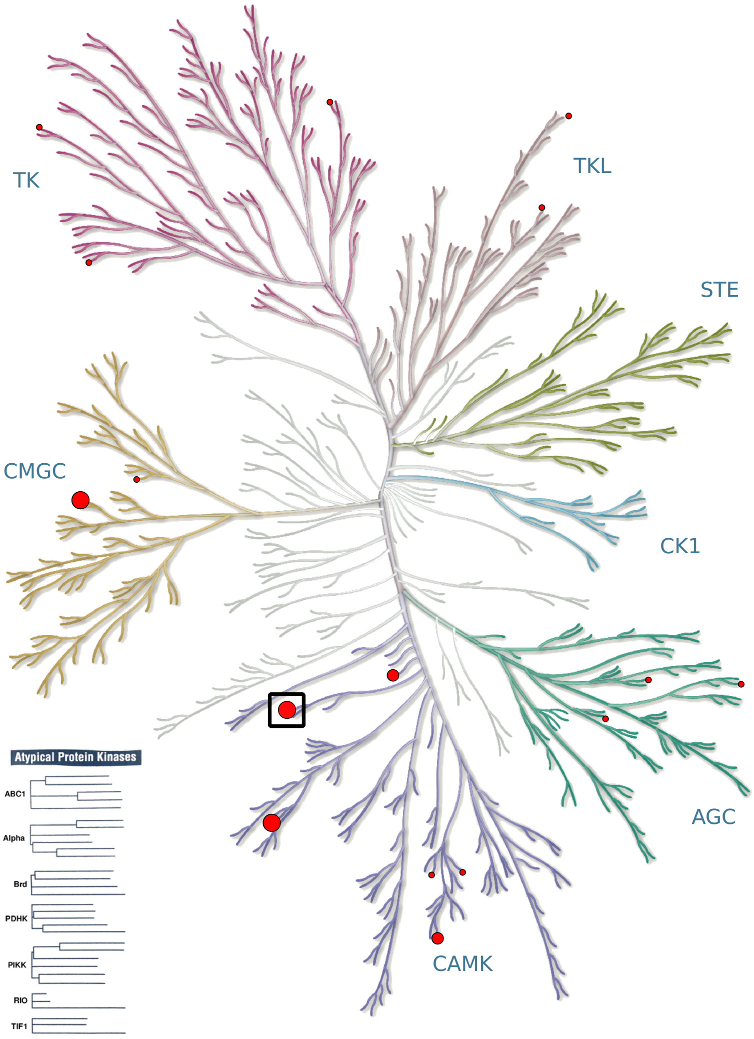


Figure 4.17: Kinase map of inhibition data for compound **102**. Circle size denotes potency category: (large: inhibition ≥70%, medium: 50%≥ inhibition <70%, small: 30%≥ inhibition <50%). Potent inhibition is observed for two of the eight groups of typical kinases, and for none of the 13 atypical families described by Manning et al. N.B. kinase inhibition data are available for 58/536 kinases visible. PIM1 kinase is highlighted. A target-labelled map can be found in Appendix 5, Figure 3. Illustration reproduced courtesy of Cell Signalling Technology, Inc. (www.cellsignal.com). Illustration produced using KinMap [466]. For comparison, see Karaman et al., SI Figure 1 [409].

Considering the kinases for which the compound is a potent inhibitor, cross-reactivity with GSK-3β is commonly observed, according to comprehensive surveys of PIM1 patent data [454, 456], with some authors suggesting this is owing to a shape-mediated interaction [461, 467]. GSK-3β has a complex network of signalling pathways, but inhibitors are in development for Alzheimer's disease [468], NSCLC and certain forms of leukaemia [469, 470]. DRAK1 has also been discussed as a potential target for head-and-neck cancers [471], suggesting the possibility of a synergistic effect with PIM1 [453].

Comparative selectivity analysis These data are difficult to interpret in isolation. Inhibition data for each kinase profiled were extracted from the ChEMBL 25 database. 167 compounds were identified which had measured activity data for 55/58 (95%) of the targets tested. We calculated selectivity metrics (see section 3.5.1) for the overall pool of compounds, and performed further analysis for the subset of these compounds with annotated PIM1 activity. We applied a threshold to all activity data of 30% inhibition, as was discussed in section 4.3.2. For the overall analysis, we highlight reference values for all compounds in the set which have progressed beyond phase I trials, to provide reference values for kinase inhibitors trialled for safety in man.

In Figure 4.18, we see that compound **102** is more selective than the majority (four fifths) of compounds in our study which had passed this critical stage in safety testing, under all metrics other than the Gini coefficient. Considering all compounds, regardless of their primary target, 28% of compounds had a lower WS(20%), along with 42% and 34% with better RS(20) and RS(5) values. This combination indicates that compound **102** potently inhibits a few targets, but that the drop-off in inhibition is substantial thereafter. The Gini coefficient was the only metric which changed substantially with the 30% inhibition threshold previously discussed, as a result of its noise-sensitivity. Without the threshold, 82% of compounds had a better Gini coefficient, which dropped to 59% when the threshold was applied to data for all compounds.

If we consider the distribution of inhibition values for our curated dataset, and for compound **102** (Figure 4.19), we find an interesting behaviour. For the dataset distribution there are many low values, but a surprisingly high proportion of very high ($\geq$95%) inhibition values. A two-sided Mann-Whitney U test, at α=0.05, fails to reject the H_0 (p value = 0.09), i.e. there is insufficient evidence to support the assertion that our data come from a different underlying distribution to the database. With

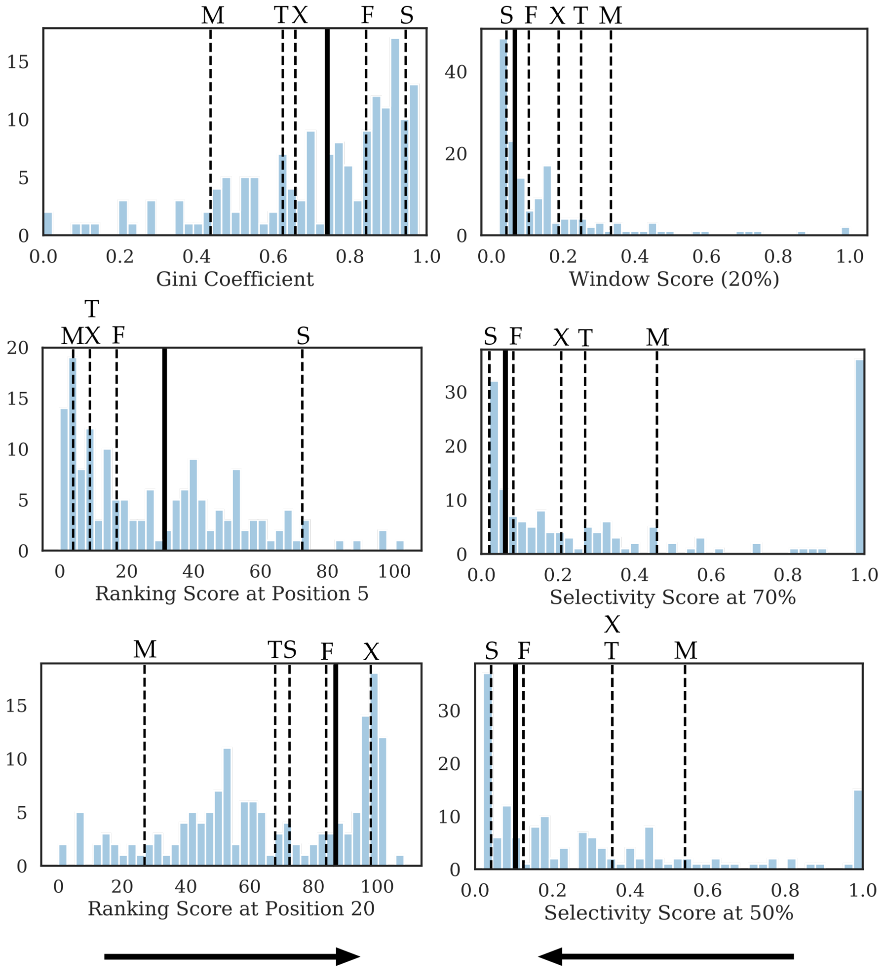


Figure 4.18: Distribution of Gini coefficient, window score, ranking scores, and selectivity scores for compound **102**, and for the 167 compounds with activity data against our panel of 55 kinases. Arrows indicate direction in which graphs show higher specificity. Compound **102** is indicated in each instances with a solid black line. Reference phase II compounds are annotated as such: 'S': Sirolimus, 'F': Fasudil, 'X': Semaxanib, 'T': Tirolimus, and 'M': Midostaurine. Vertical stacking of letters indicates that the reference compounds were indistinguishable with that metric. Under all scoring methods bar one, compound **136** had better specificity than four of five reference compounds which had passed phase I trials in man.

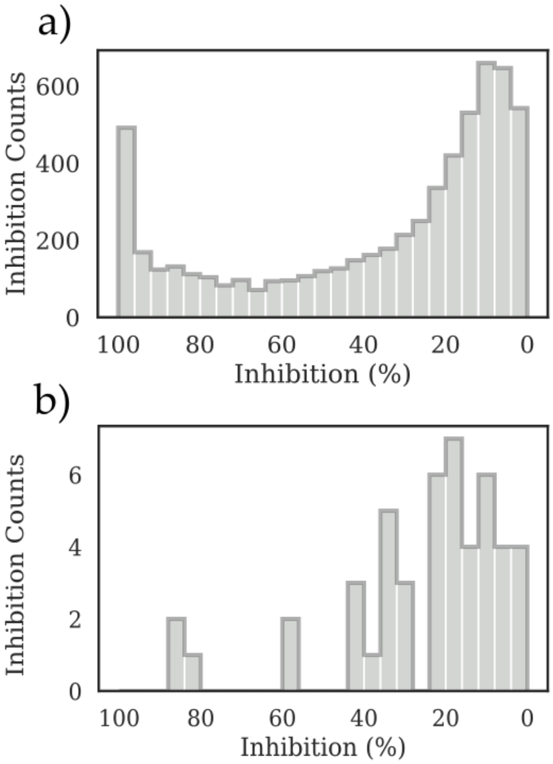


Figure 4.19: a) Distribution of inhibition values for all pairs of compounds (*n*=167) and targets (*n*=55) in our training set. Inhibition values precisely equal to zero (*n*=1807) have been removed, for emphasis. b) Distribution of inhibition values for compound **102** for all targets.

either interpretation, however, **102** is comparatively selective, with good separation between high-inhibition targets, including PIM1, versus the rest of the kinases assayed.

PIM1 Inhibitors With a more specific focus on the selectivity of PIM1 inhibitors, we isolated the subset of our dataset annotated as active against PIM1 at 10 µM (*n*=71). Of these, six have both a higher percentage inhibition of PIM1 and selectivity under the Gini, RS(5), and RS(20), 8 under WS(20), and 11 under WS(5) (Figure 4.20) than compound **102**. In summary, compound **102** has amongst the best overall selectivity profiles of the PIM1 inhibitors retrieved from ChEMBL under most metrics. In Figure 4.20 we also see the distribution of pairwise similarities for all other compounds profiled, against compound **102**, under FDFP8 and ECFP6. There is no clear relationship between either of these similarity metrics and the specificity values obtained, likely owing to the low similarity between compound **102** and those for which we could find adequate selectivity data (ECFP6-Dice: mean 0.13±0.11, max 0.22, FDFP8-Dice: 0.32±0.11, max 0.60). As such, even though the compounds profiled are much-closer in shape-space than under a topological approach, we are still relatively far from our confidence threshold. The most similar compound under FDFP8 has a ΔGini coefficient of 0.04, ΔS(50) of 0, and ΔRS(20) of 4, compared to the ECFP6 most similar with 0.13, 0, and 13, respectively.

Given these relatively low similarity values, we instead compare compounds based on their potency and selectivity. Considering the most selective compound which has at least as potent an inhibitory effect as compound **102**, compound **138** (96% inhibition at 10 µM, Gini 0.91, S50 0.06), and the most potent PIM1 inhibitor, compound **139** (100% inhibition at 10 µM, Gini 0.62, S50 0.35), we see pronounced differences in the kinases inhibited. We will refer only to those kinases inhibited at the moderate and potent levels, for brevity. Staurosporine is an equally-powerful inhibitor of PIM1 as compound **139**, but with a significantly poorer selectivity profile.

The most selective highly-active compound, **138** (Figure 4.21), has increased inhibition of the protein kinase G (PKG) and Lyn kinases. Inhibition of PKG is associated with an increased risk of cardiac hypertrophy [472] and heart disease. Lyn inhibition has been noted to result in slowed growth of acute myeloid leukaemia cell lines, and similar effects in other leukaemic cell lines. Additionally, Lyn-specific inhibitors have been observed to disrupt certain kinds of acquired resistance to the drug imatinib [473]. The effect of Lyn-specific inhibitors on the immune system remains unclear, but Lyn-knockout mice develop more severe and persistent asthma than control animals. Compound **138** shows no significant inhibition of the targets GSK3β, DRAK-1, CHK1, CK1γ1 and Rsk1 kinases.

By contrast, compound **139** (Figure 4.22) inhibits NEK2, SRPK1, CK2a2, ALK, IRAK4, MAP2K1, PAK2, PLK3, PKG1, ROCK1, AurA, MAPKAPK1, MAPKAPK5 and FRAP kinases, in addition to PIM1 and GSK3β. All three compounds share some level of activity against the GSK3β and PKG1 kinases, and two out of three inhibit Lyn, SRPK1, or CK2a2.

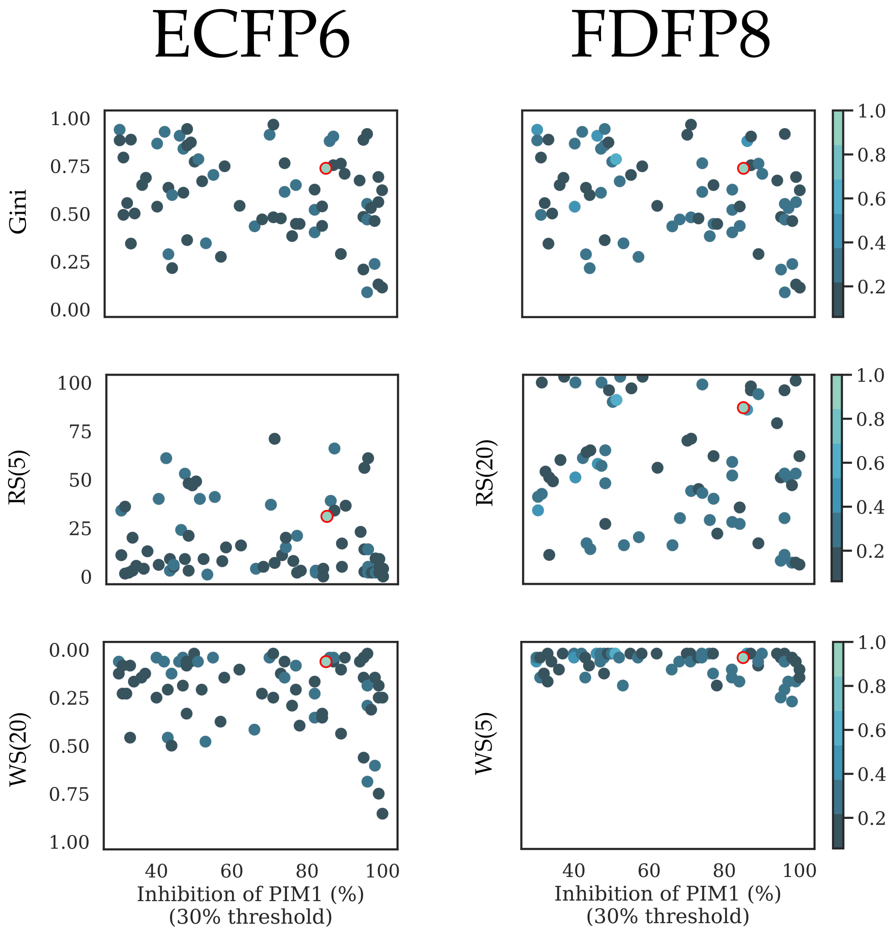


Figure 4.20: Gini coefficient, window score, ranking scores, and selectivity scores for compound **102** and the 71 compounds with activity data against our panel of 55 kinases, and a greater than 30% inhibition of PIM1 at 10 µM. Compound **137** is indicated in each case with a red circle. Graphs in the left-hand column are coloured according to the similarity of each compound to **137** under ECFP6, using Dice similarity. The right-hand column is coloured by similarity under FDFP8, also employing Dice similarity. Each graph is oriented such that the upper part of the y-axis represents better specificity. Under every specificity metric other than WS(5), compound **137** is amongst the top 10% of PIM1 inhibitors assessed when considering specificity and activity. No clear relationship between ECFP6 or FDFP8 similarity and specificity can be established. Average and maximum similarity to other compounds profiled under ECFP6: 0.13±0.11, 0.22. Average similarity to other compounds profiled under FDFP8: 0.32±0.11, 0.60.

4.4.3 Crystallographic study

To gain another perspective on the matter, and to determine whether the relatively low similarity values to well-characterised antagonists are owing to a novel binding mode, we obtained a crystal structure for compound **102** from a fee-for-service provider (SARomics Biostructures AB, Lund, Sweden). The PIM1 ligand complex was solved to 1.8Å, with clear density for the inhibitor, and full occupancy in the substrate-binding site. A list of crystallographic parameters is provided in Appendix section 5, Table 2.

PIM1 Crystal Structure Analysis Having obtained a crystal structure for **102** in complex with PIM1, we gathered existing structures for comparison using the RCSB PDB API (*n*=155). We cleaned all structures using the PDBFixer [377] script. Every remaining valid PIM1 crystal structure was aligned to our obtained structure using a sequence-weighted structural alignment routine (see subsection 3.5.2, PIM1 - Inhibitor crystal structure) (sequence identity: 96.1%±16.3, *n*=141), saving the transformed co-ordinates into new PDB files, and recording the RMSD of each transformed structure with regard the template (Average: 1.00±0.49Å) . This approach was chosen as the sequence length of the human PIM1 constructs crystallised vary substantially.

The crystal poses of compound **102**, and of the 133 ligands previously crystallised, are included in Figure 4.24. We see that compound **102** binds in the same binding pocket as all other known inhibitors, in the hinge region of the PIM1 kinase. The compound has a specific interaction with Lys67, which is highly-conserved across known inhibitors [474, 475], via a charge-assisted hydrogen bond with the furan moiety. For all solved structures of PIM1-inhibitor complexes, there are interactions with either or both of Lys67 and Glu121, accepting and donating hydrogen bonds, respectively. In addition, there are multiple potential hydrophobic interactions between the central scaffold and buried residues in the binding pocket (Leu44, Ala65, Leu120, Leu174, Ile185). We see preservation of the known salt-bridge interaction between Lys67 and Asp186 [476, 477].

Compound **102** has a relatively good average B-factor [478], considering the high resolution of the solved crystal structure, with an average of 34.3 Å^2 across the molecule. The central scaffold of the molecule is especially stable, with a good B-factor, full occupancy, and no predicted 'bad contacts', using the Maestro [479] ligand interaction tool. The two terminal regions, a furan and a benzene, are somewhat less stable, and

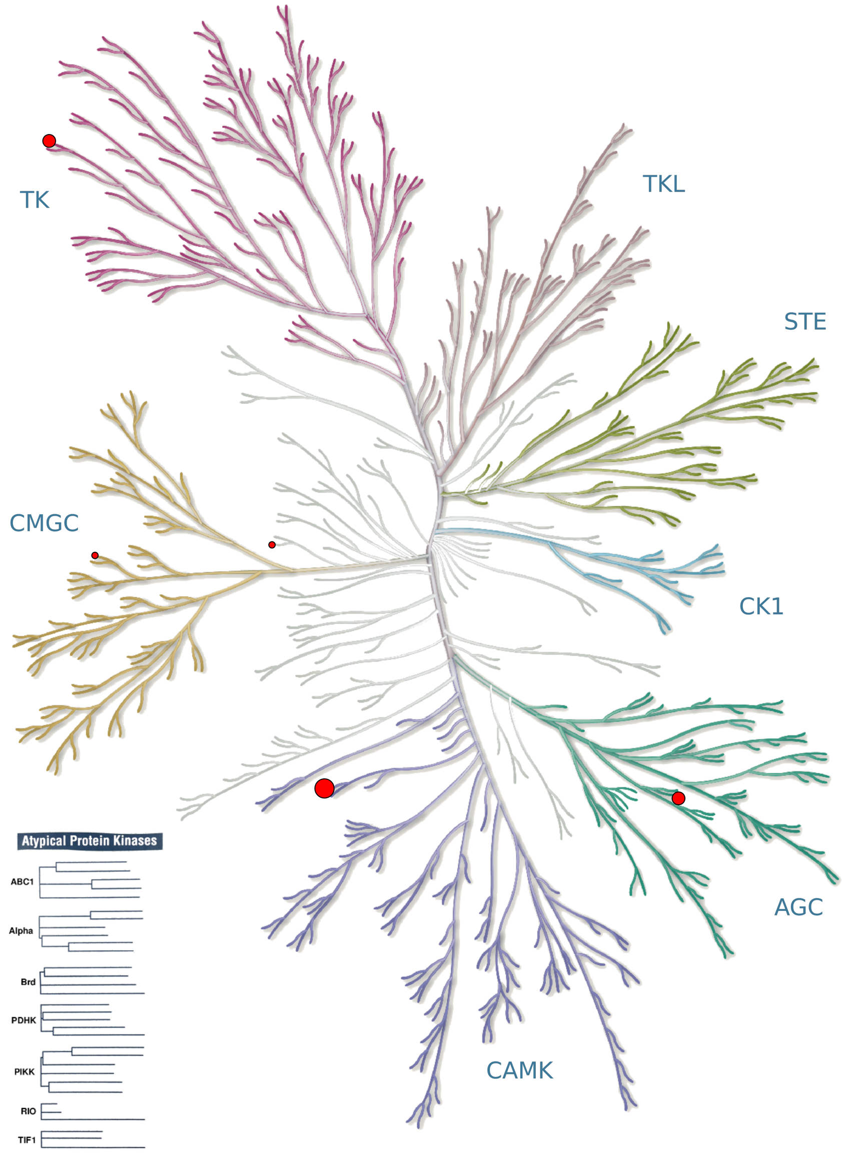


Figure 4.21: Kinase map of inhibition data for compound **138**. Circle size denotes potency category: (large: inhibition ≥70%, medium: 50%≥ inhibition <70%, small: 30%≥ inhibition <50%). Potent inhibition is observed for six of the eight groups of typical kinases, and for one of the 13 atypical families described by Manning et al. N.B. kinase inhibition data are available for 58/536 kinases visible. Illustration reproduced courtesy of Cell Signalling Technology, Inc. (www.cellsignal.com)

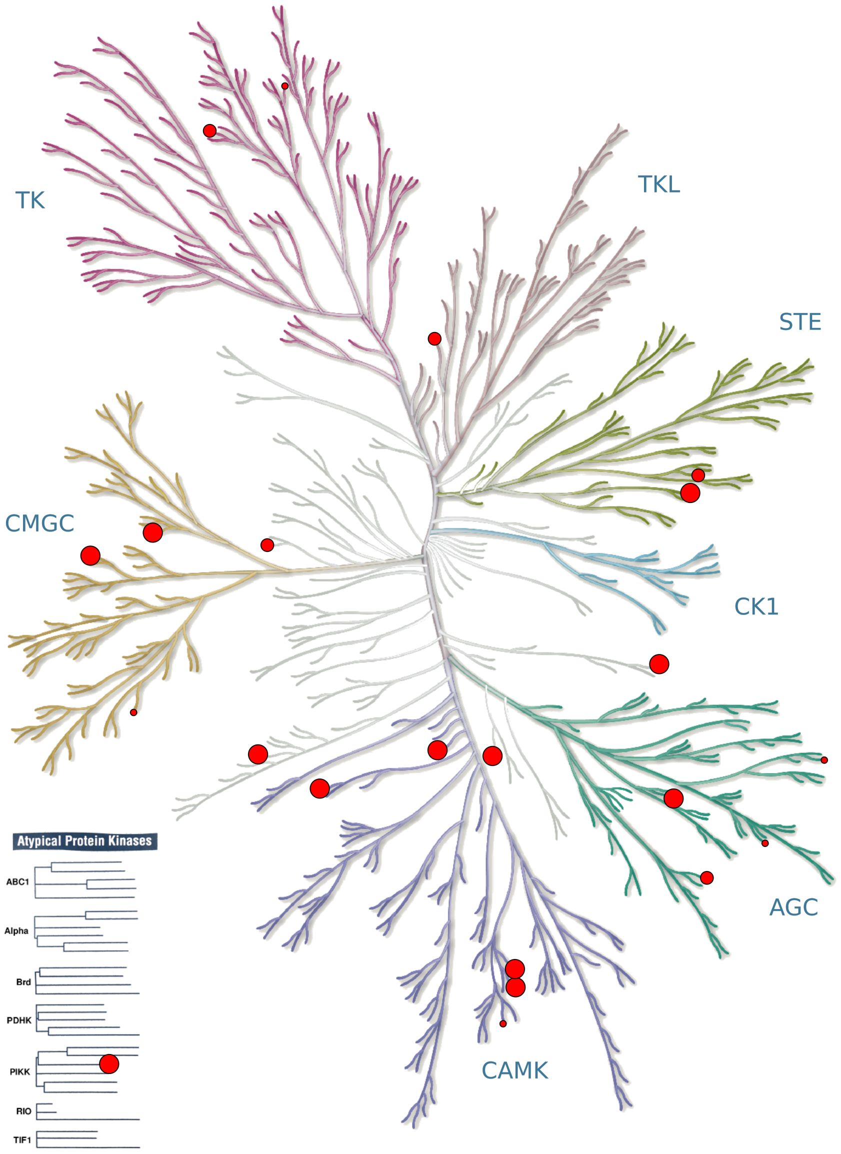
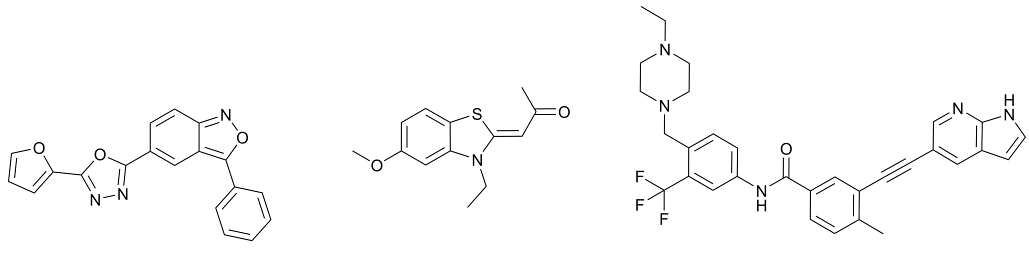


Figure 4.22: Kinase map of inhibition data for compound **139**. Circle size denotes potency category: (large: inhibition ≥70%, medium: 50%≥ inhibition <70%, small: 30%≥ inhibition <50%). Potent inhibition is observed for two of the eight groups of typical kinases, and for none of the 13 atypical families described by Manning et al. N.B. kinase inhibition data are available for 58/536 kinases visible. PIM1 kinase is highlighted. Illustration reproduced courtesy of Cell Signalling Technology, Inc. (www.cellsignal.com)



Compound 102 Compound 137 Compound 138

Figure 4.23: Structures of compound **102**, the most selective compound profiled, compound **138**, and the most potent inhibitor, compound **139**.

solvent-exposed. These moieties have a high degree of flexibility, and might be suitable for improvement through medicinal chemical efforts.

The structure of the PIM1 kinase, in complex with all of the inhibitors for which a crystal structure is publicly available, seems relatively stable. The small loop region, containing amino acids 43:50 is comparatively unstable, as can be seen in Figure 4.25 and Figure 4.26.

PIM1 Crystal Structure FDFP Analysis We utilised our FDFP approach to analyse the ligand and protein components independently for our set of crystallographic data, and compared its performance on the first task with ROCS shape and combination approaches. For each PDB file, we extract the ligand in its crystal conformation, generate a set of conformers of that ligand, and store these alongside the apo-protein. For the analysis of ligand crystal-derived and energetically-minimised conformations, we define four sets of pairwise comparisons; 'Self', where we consider the maximum similarity of a given ligand's crystal conformation to those generated by the maximisation routine, 'Crystal-Crystal', where we compare the similarity of the crystal pose of compound **102** to the crystal conformations of all other ligands, 'Crystal-Minimised', which assesses the maximum pairwise similarity of crystal structure of compound **102** against each set of conformers for the remaining molecules, and, lastly, 'Minimised-minimised', which replaces the pose obtained for compound **102** with the set of its generated conformers. Pocket comparisons are performed using FDFP8 only (see subsection 3.5.2). As each method profiled has a different innate scale of similarity values (See Figure 4.4), comparisons are primarily in terms of relative shifts in the distributions, their mean values, and standard deviations.

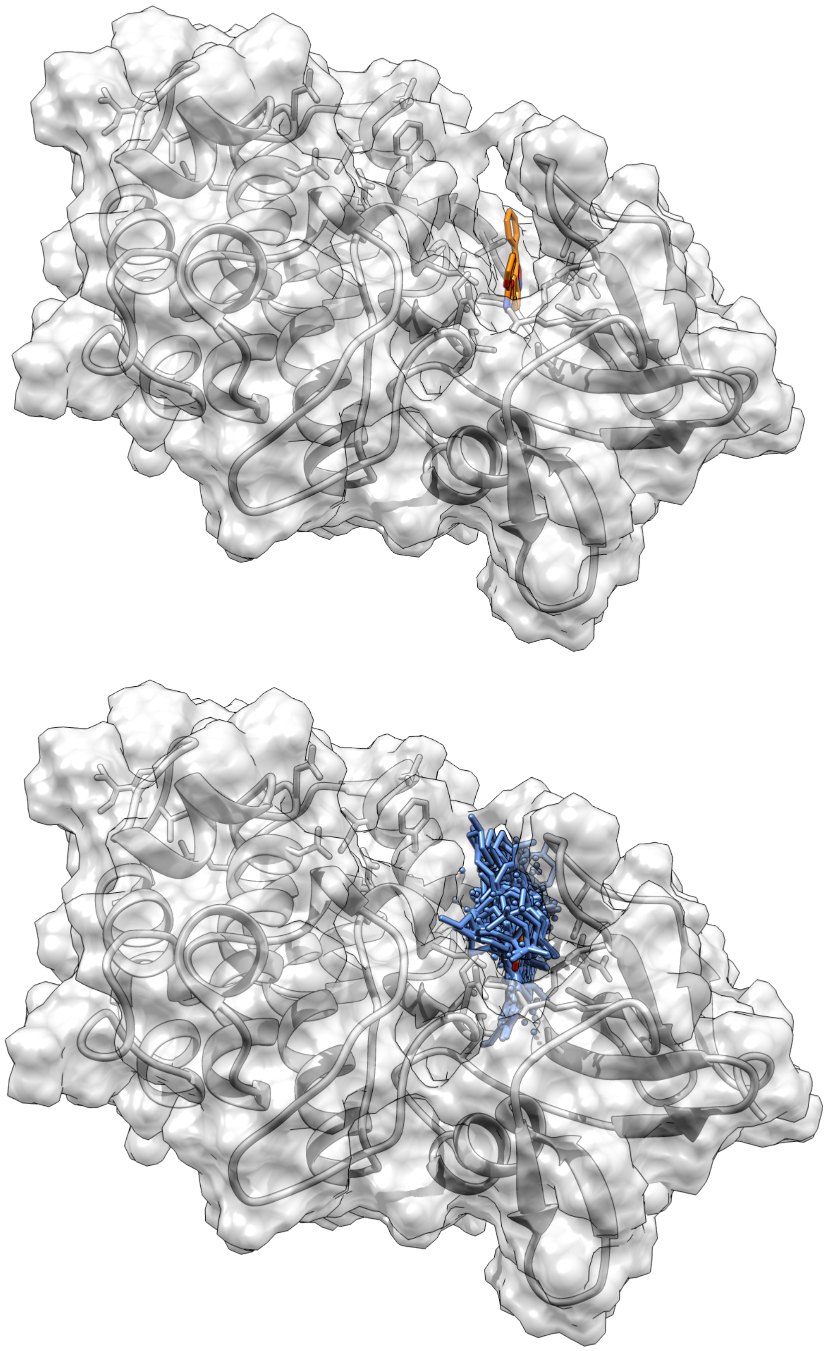


Figure 4.24: Visualisation of PIM1 in complex with compound **102**. Top: Ligand localised in known binding pocket. Bottom: Overlay of ligands from collection of superimposed PIM1 crystal structures extracted from the RCSB PDB (*n*=133)

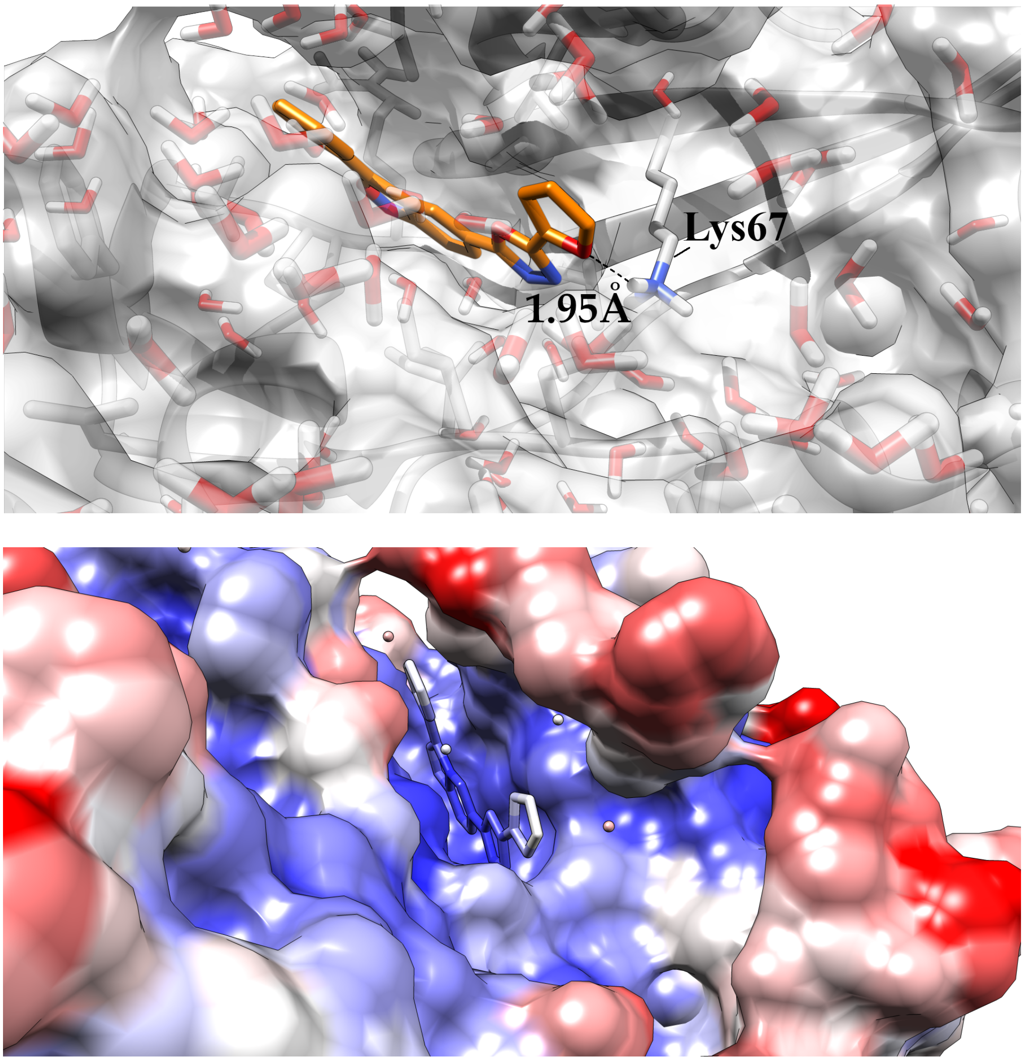


Figure 4.25: Top: Compound **102** has a hydrogen-bond interaction with Lysine-67, known to be important for PIM1 inhibitor activity, as well as several hydrophobic interactions with the binding pocket. Crystal waters shown in transparency. Bottom: Visualisation of B-factors for crystal structure of **102** in complex with PIM1. Colourisation: Blue to white, then red, at 0, 40, and 100 Å^2 [478]. The central scaffold of the molecule is stable, with no crystallographic waters noted in that deeper part of the pocket. The furan and benzene termini are less stable, and solvent-exposed.

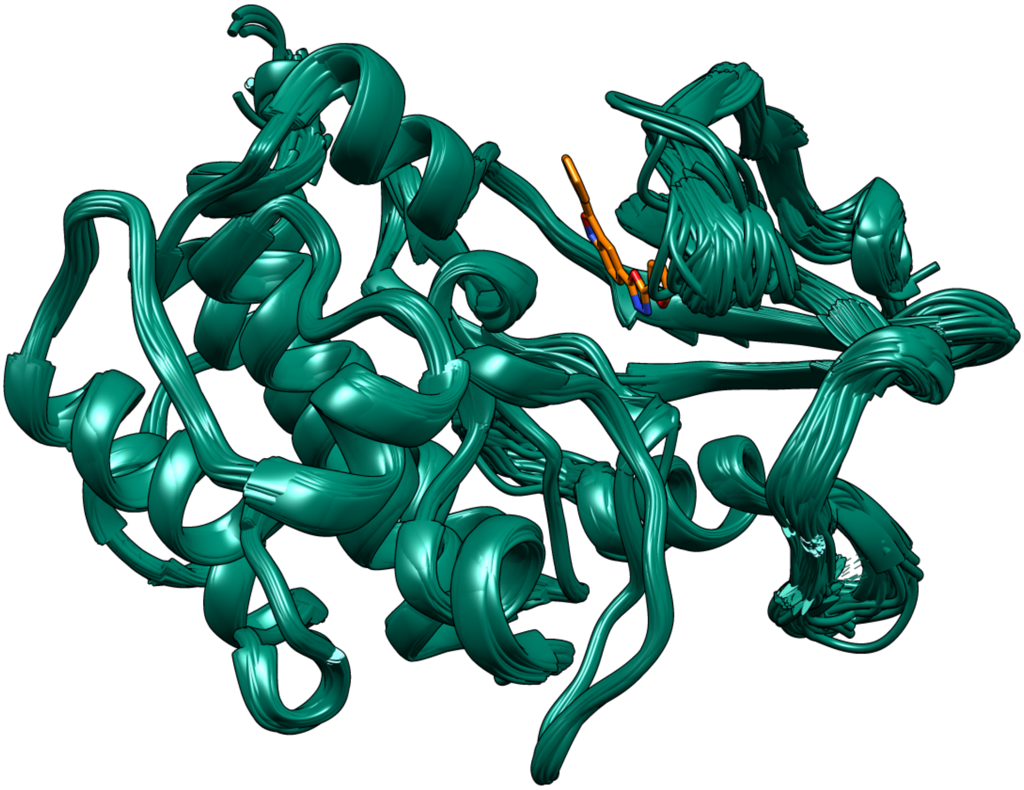


Figure 4.26: Structural overlap of the protein components of 133 processed crystal structures of PIM1 downloaded from the RCSB PDB, superimposed on our crystal structure with compound **102**. Overall, the secondary and tertiary structure of the protein is relatively stable. The largest variation near the substrate binding pocket is in a small loop region. The relative instability of this region can also be seen in the B-factor map in Figure 4.25. In some ligand-bound conformations, this loop region partially obscures, but does not seem to occlude, the position of the furan moiety, which forms an important interaction with Lys67.

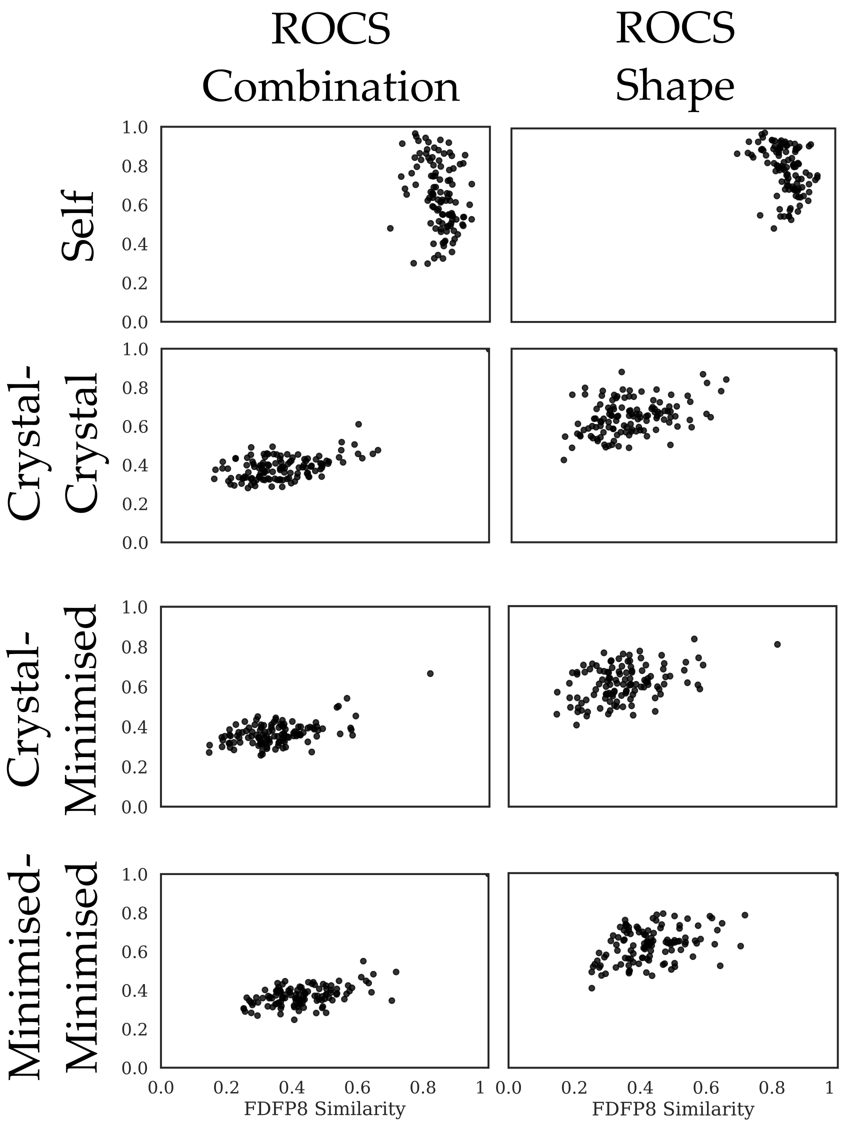


Figure 4.27: Similarity under FDFP8, ROCS Combination and ROCS shape for ligands with crystal structures for PIM1 Kinase. 'Self' describes the maximal pairwise similarity between generated conformers for a given ligand, and its solved crystal structure. 'Crystal-crystal' describes the pairwise similarity between the crystal poses of each ligand, compared to that of compound **102**. 'Crystal-Minimised' describes the pairwise similarity between the minimised conformations of each ligand, compared to the crystal structure pose of compound **102** and 'Minimised-Minimised' similarly, replacing the crystal conformation with the corresponding set of conformers.

Table 4.23: Distribution of similarity values generated for four cases. Averages and standard deviations are provided, as well as the mean value of the five-highest similarity values in each case (in parentheses), as a measure of the tailing in the distributions. Similarity is highest for the 'self' case, and lowest when comparing crystal and minimised conformers. Definitions are as per subsection 3.5.2, and Figure 4.27

	FDFP8	**ROCS Comb**	**ROCS Shape**
Self	0.84±0.04 (0.93)	0.64±0.17 (0.94)	0.78±0.1 (0.95)
Crystal-Crystal	0.37±0.12 (0.70)	0.39±0.08 (0.63)	0.65±0.09 (0.88)
Crystal-Minimised	0.35±0.1 (0.63)	0.36±0.05 (0.53)	0.61±0.09 (0.79)
Minimised-Minimised	0.43±0.11 (0.74)	0.37±0.07 (0.60)	0.63±0.09 (0.82)

In the 'self' case, we see high average similarities under FDFP, and ROCS Shape, with somewhat worse results for ROCS combination (Table 4.23, Figure 4.27). Notably, the standard deviation varies substantially between methods. In Figure 4.28, the conformation for compound **102** obtained from crystallisation is compared with those generated in the diversity-oriented conformer generation. While the central, relatively rigid, adenosine-mimetic scaffold is well-captured, the termini show considerable divergence, and none of the conformers generated has the same essentially-flat pose as that obtained from the crystal structure, which has been observed in several solved structures of PIM1 inhibitors [475, 476].

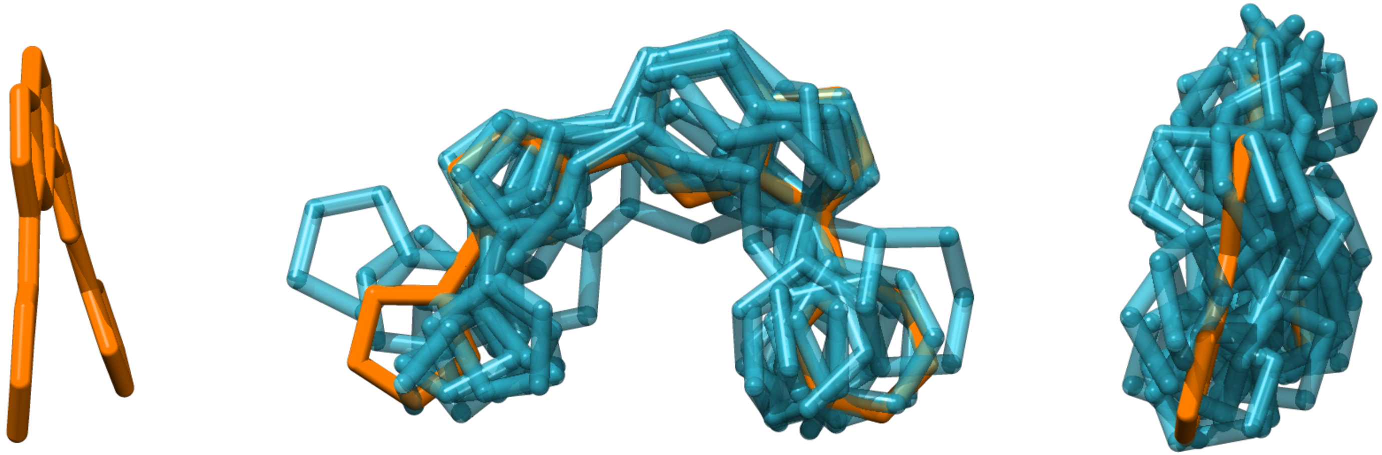


Figure 4.28: Left: pose of compound **102** obtained from the crystal structure. Centre, right: Alignment of generated conformers of compound **102** to the conformation obtained from the crystal structure. Diverse conformers were generated using the procedure described in subsection 3.1.4, Conformer generation. The relatively flat conformation obtained from the crystal structure is not well captured by the minimisation strategy, especially in the relatively flexible termini.

This relationship changes somewhat for the other cases. In each case, the distribution of similarities is significantly down-shifted in the 'crystal-crystal' and 'crystal-minimised' configuration, reflected in the mean of the five maximal values. FDFP8 retrieves some compounds above the similarity cutoff threshold (see Table 4.21) earlier defined in each case, which is not true for the ROCS combination approach. As discussed earlier, probability estimation for ROCS shape is difficult, given the right-shifting and limited range of its distribution. Average similarity of the FDFP8 method increases somewhat when considering the minimised-minimised set, an effect which is not observed for ROCS shape or combination methods.

Having considered shape-similarity from a ligand perspective, we then considered the binding pockets themselves in terms of their shape fingerprints. Protein FDFP8 fingerprints were calculated for all 134 crystal structures, and folded to 512 bits. We then calculated pairwise similarities for each of these against the crystal structure obtained for compound **102**. The average similarity obtained was 0.26±0.06 (Figure 4.30) . As we have no empirical similarity cutoff for this approach, we instead examine the order imposed on the set of protein crystal structures when ranked by pocket similarity.

Checking each structure manually, we found that the relevant compounds were superficially similar initially, morphing into larger molecules growing upwards out of the cleft, and finally into small, adenosine-like inhibitors. In general, the ordering seems relatively intuitive, given its progression from similar compounds, to similar compounds interacting with more residues, to those interacting with fewer, although alternative distance metrics would handle the absence and presence of features differently. On a quantitative basis, there is slight correlation between protein FDFP8 and ligand FDFP8 (Kendall-Tau, 0.17, p = 0.02), and also with ROCS Shape (0.13, 0.03). No statistically-significant correlation was noted with ROCS combination similarity scores.

In Figure 4.29, we see the most similar ligand under ligand (compound **102a**, similarity = 0.70) and protein **102b** FDFP dice (0.43) similarity to the bound conformation of compound **102** and its binding pocket, respectively. In both cases, molecules sharing many generic features are retrieved, with similar central scaffolds, and similarly-sized decorations. Both molecules extend towards the Lys-67 residue, and conserve a similar binding pose, with more variation in the terminus farther from the identified interaction.

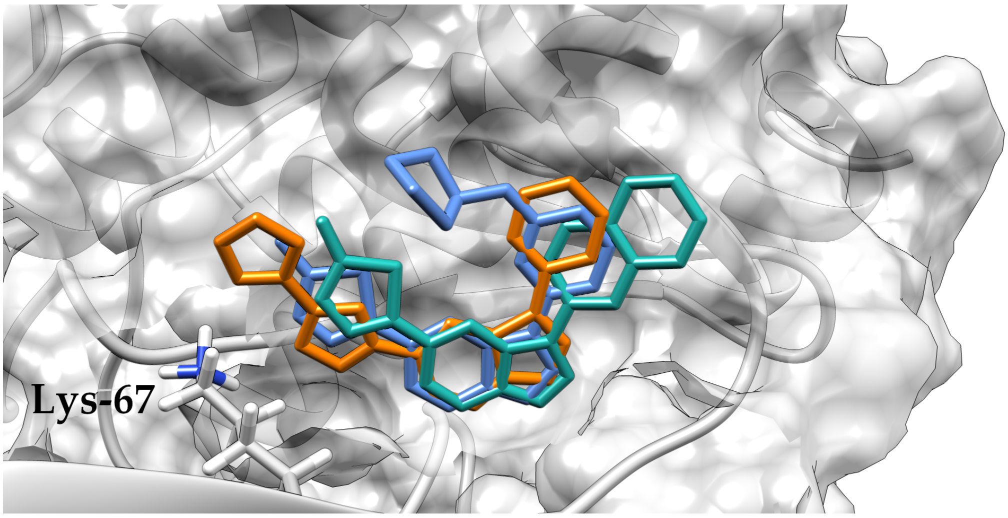


Figure 4.29: Superimposition of the ligand (compound **102a**) with the highest similarity under FDFP8 to compound **102** (PDB-ID: 4WSY, light green) and of the ligand (compound **102b**) from the binding site with the highest similarity under FDFP8 to the binding site of compound **102** identified during the crystallographic study (PDB-I: 4WT6, blue)

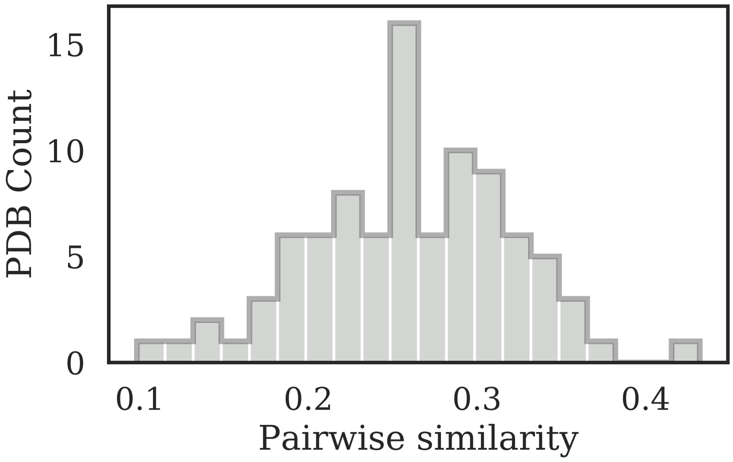


Figure 4.30: Distribution of Dice similarity values of the FDFP8-512 bit representations for each of the 133 previously-solved structures of PIM1 in complex with an inhibitor against the equivalent representation of the crystal structure obtained with compound **102**. Although values are lower than in the ligand case, there is good subjective agreement as can be seen in Figure 4.29.

4.4.4 Discussion

We identified five inhibitors of PIM1 kinase with at least 30% inhibition at 10 µM, further characterised as four high-nanomolar or low-micromolar inhibitors, using the FDFP approach, of a total of 17 compounds profiled. Of these, all were located in the above-threshold (see Table 4.21) region, which contained seven compounds. As such, the hit rate for our above-threshold picks is either 71%, or 29% if we take the whole set. Each of these compounds represents a novel BMS (see Table 4.20), and three novel GMS in total, indicating that the retrieved compounds represent novel chemistry for this target. This compares to a hit rate of approximately 0.3% for this target by HTS (by Vertex pharmaceuticals) [474], 1% for a virtual screening [480], 4% for docking [474], and 25% for a multi-level virtual screening approach, incorporating similarity-searching, machine-learning, and docking [481], of which 25% had novel scaffolds.

The most potent inhibitor identified after IC_{50} determination was chosen for further profiling, and found to be relatively selective compared to published compounds for which sufficent data could be obtained. Interestingly, we saw potential evidence of reporting bias in our selectivity dataset (see Figure 4.19), with a significant peak for very potent (>95% inhibition at 10 µM) inhibitors of all 55 kinases profiled. It is unclear whether this skew in the reporting of very highly-active compounds is some fundamental property of kinase inhibitors, or reflects practices surrounding compound activity data disclosure. While target specificity was not explicitly selected for (i.e., we did not limit the pool of templates described in Table 3.2 to those for which a certain selectivity threshold was met, for example), this is a promising outcome. Although it has been proposed that shape similarity can more directly encode some specificity-determining features than pharmacophoric approaches [482], the evidential basis for this is still somewhat lacking. A common stance [73, 132] is that shape complementarity between a ligand and target pockets alone is likely necessary, but insufficient, to determine specificity. To a certain extent, the established ability of shape-based methods to find novel scaffolds which are active against a particular target suggests that they capture the features which are a prerequisite to binding; however, as identity of the atoms in a molecule, and the pattern of these constituting a conventional pharmacophore, determines the geometry of a molecule, it is difficult, and to a certain extent incoherent, to define which component best selects for specificity.

Keeping in mind the diversity of ligands retrieved over the set of seven targets (see Table 4.20), two simple, and somewhat opposing, models suggest themselves for further investigation. In the first, pharmacophoric features serve as the primary 'interaction surface' determining binding, and ligand geometry provides sufficient structure to ensure these are in the appropriate geometric configuration. In the second, ligand geometry predominates, and the pharmacophoric features add an additional determinant [254]. In favour of the first model are the energetics of ligand binding, with notably higher bond energies for specific interactions. For the latter, we note that scaffold hopping, by necessity, requires substantial alterations in the identity or configuration of atoms within the central scaffold. One might assess the relative contribution of each by a broader cross-comparison of molecules, known to be active against a certain target, under pharmacophoric and shape approaches. The issue of multiple binding sites renders this somewhat difficult, but here the kinases could provide a useful platform for investigating this relationship, given their canonical binding site, abundance of similar structures, and availability of substantial ligand-target data. As such, it seems a useful starting point for a broader investigation into the role of shape in determining specificity. More specifically for PIM1, the large dataset of kinase specificity assembled here could help guide the development of more specific inhibitors by both positive and negative design [483], selecting compounds closer to known selective inhibitors or potent on-target compounds, and further away in shape space from known inhibitors of off-target kinases.

An additional benefit of kinases as a platform for investigating specificity-determining factors is the relative abundance of crystallographic data. We found that the profiled compound binds in the canonical binding site, in line with earlier findings that shape-similar compounds share similar binding sites and poses [484], although unsurprising for a kinase inhibitor. In our comparative crystallographic study, we observed some interesting properties of existing shape-based approaches as applied to structural data. The narrow dispersion of FDFP8 (see Table 4.23) , and the comparatively wider-dispersion of the other approaches, supports the argument, discussed in section 4.1.1, that FDFP takes a somewhat orthogonal approach to the ROCS family of alignment-based approaches, and varies less with the conformational disposition of the query and template molecules. This might serve to explain the relatively low similarity values obtained by ROCS combination on the self-similarity task (when compared to the distributions observed in Figure 4.4 for example, and which are also in many cases lower than the threshold obtained from our probabilistic model), as we found that our conformer generation routine struggled to approximate the bound conformation in this case (see Figure 4.28).

Supporting the hypothesis that FDFP8 is somewhat orthogonal to the ROCS approaches, the Kendall-Tau correlation coefficients between the maximum similarity and minimum RMSD for each compound are 0.09 (p=0.36) , -0.88 (p=0.007), and -0.92 (p=0.003) for FDFP8, ROCS Combination, and ROCS Shape, respectively, indicating that the quality of both ROCS similarity measures is extremely dependent on the accuracy of the conformers generated, whereas FDFP is to a large extent independent of this, at least in this case. The importance of replicating crystal structure conformations is well known for ROCS, and a primary reason for the relatively large conformational ensembles typically employed [385]. The reason for the mismatch between ROCS combination and shape is unclear, as in this case the electrostatic features are identical, but unfortunately the specifics of the ROCS implementations as they currently stand are somewhat unclear. Interestingly, we see no substantial increase in similarity for either ROCS method when considering the minimised-minimised case (in which we compare generated conformers of compound **102** to those generated for each of the other solved ligands). We are not entirely clear why this should be the case as of yet, given that the ROCS combination method produced superficially similar compounds when utilised on the ChEMBL set (see Figure 4.15). Repeating this analysis for a broader set of crystal structures, covering a wider swathe of target space, might serve to shed some light on the peculiarities of each method.

4.5 Back-translation study

Summary We observed qualitative agreement of the ligand and protein FDFP binding descriptions in our PIM1 case study, and a modest correlation. To better understand the relationship, we sought a means of describing this complex relationship through statistical approximation. Having tried simple approaches, such as FFN with binary representations, we determined that our model required a higher descriptive power. As such, we considered this as a sequence transduction problem, which can be thought of as training models to generate ligand 'answers' to protein 'questions', or as translating a protein pocket into a ligand complement. Utilising a large database of crystallographic structures, we found an architecture which managed to generate reasonable reproductions of ligand shape fingerprints, based on analysis of the binding pockets, for approximately two-thirds of cases, and excellent facsimiles for one fifth of targets.

4.5.1 Results

scPDB analysis We utilised the scPDB database (2017 release) of ligand-protein crystallographic data for the purposes of our proof-of-concept study. For each of the valid complexes in that database (n = 14,556), we generated FDFP fingerprints for both the ligand and its corresponding protein pocket. For descriptive purposes, we calculated the pairwise Dice similarity between all pairs of protein FDFP6, ligand FDFP6 (bound and minimised) and ECFP6 fingerprints, as reference values (see Figure 4.31).

Overall, we find that most ligands in the database are quite dissimilar to one another, with a low ECFP6 Dice similarity (0.23±0.18). As such, this database represents a broad sampling of ligand chemical space. Interestingly, we find that bound-state FDFP6 distributions show higher overall similarity (0.49±0.14) than minimised-state (0.39±0.15), perhaps reflecting their higher overall energy [485, 486]. Protein pairwise similarity more closely resembles the distribution for minimised ligands, although it does not display the same 'long-tail' behaviour: as such, there are no 'exact matches' in that set (0.33±0.09).

Model selection The protein and bound-state binary fingerprints were transformed into text-based representations (see subsection 3.6.1). Memory limitations in our training setup did not permit consideration of the FDFP8 fingerprint, as the necessary output length overflowed the

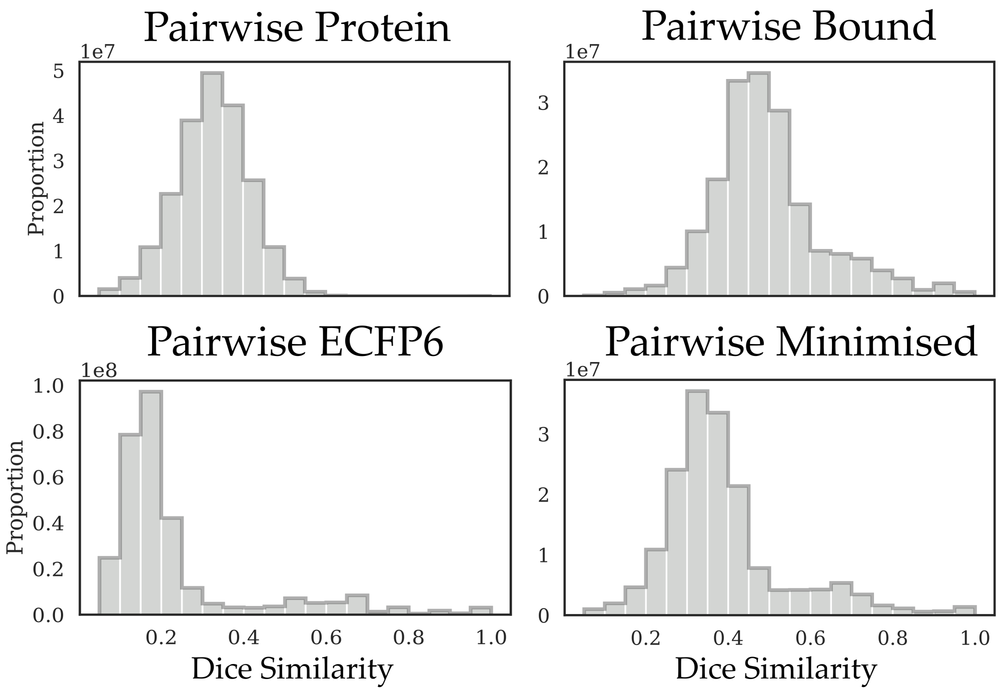


Figure 4.31: Distributions of pairwise Dice similarity for the protein and ligand components of the scPDB database. As well as 'pairwise protein' (pairwise protein FDFP6), we have 'pairwise bound' (pairwise bound-conformation ligand FDFP6), 'pairwise ECFP6', and 'pairwise minimised' (pairwise similarity of extracted, minimised ligands). As per the ECFP6 graph, we see that most of the bounds ligands are dissimilar under a conventional approach. With pairwise bound and minimised graphs, we see that bound conformations are, on average, more similar to each other than minimised conformations. With pairwise protein, we see that overall average similarity between protein pockets is relatively low, more closely resembling the distribution of minimised than bound ligands, but absent the 'long tail'.

memory buffer. Therefore, we proceeded with the FDFP6 512-bit representation for this proof-of-concept study. Models were derived from two publicly-available code repositories; an implementation of the original AIAYN transformer architecture, and the OpenNMT python package.

For a wide-scan of available sequence transduction models, we utilised the OpenNMT package. In Table 4.24, we found that the three best models were three transformer-architecture approaches. The relatively simple RNN methods, augmented with attention layers, perform well [487], followed finally by some of the larger transformer architectures

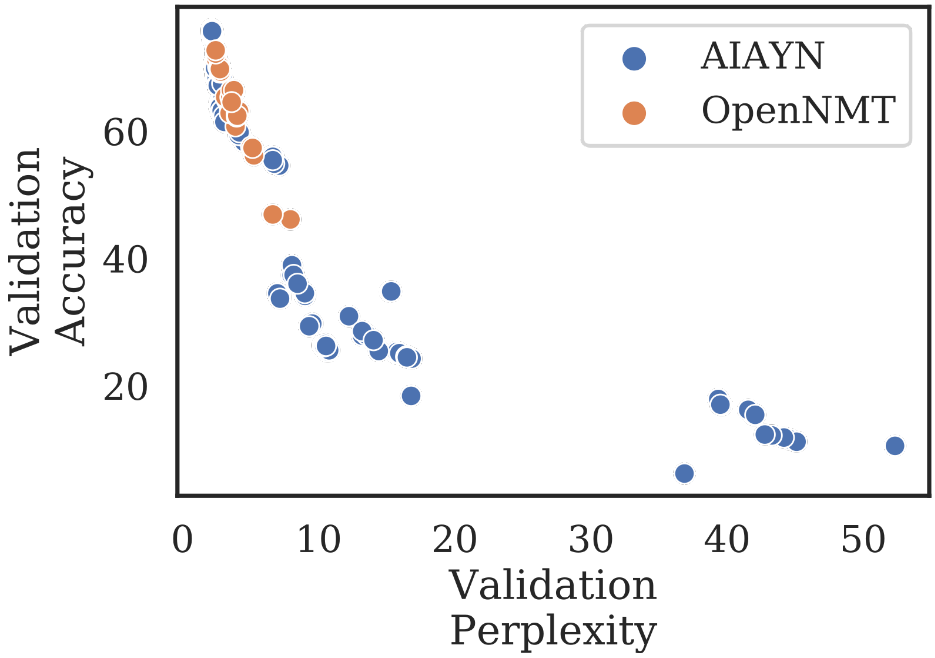


Figure 4.32: Distribution of validation perplexity and accuracy values for the models considered, coloured by their origin package. Overall, less variation was observed with default parameters for OpenNMT models than for our hyperparameter optimisation study, primarily owing to the influence of the optimiser chosen.

and convolutional approaches. Simply reducing the complexity and parameter count resulted in a substantial improvement in the validation perplexity achieved with the Google-Big architectures. Given that the primary difference between the top- and bottom-ranked transformer models are their layer widths and overall complexity, we considered this aspect in our hyperparameter optimisation with the AIAYN implementation.

In our assessment of the effect of hyperparameters on transformer performance, we found that, in this case, the choice of optimiser was by far the largest determining factor [488]. Higher model complexity resulted in poor performance for the optimisers other than Adam, which proved comparably resilient to parameter modification. Model width and depth had no significant effect on the outcomes. We chose the most parsimonious, lowest validation-perplexity model (AIAYN, Adam optimiser, four layers deep, 512 width, eight attention heads) for further study.

Table 4.24: Results of the OpenNMT model screen. Results are sorted by ascending validation perplexity, where a higher value indicates that the model is more uncertain about its predictions. For model definitions, please see section 3.6.2

Model Name	Validation Perplexity	Validation Accuracy
MINI-Transformer	2.42	72.59
MED-Transformer	2.43	72.23
MEDWIDE-Transformer	2.43	72.26
GRU	2.70	70.16
LSTM	2.74	69.66
OpenNMT-Google Big-Shallow	3.47	63.77
OpenNMT-Google	3.51	65.33
CNN-9	3.56	66.44
OpenNMT-Google Shallow	3.89	60.73
CNN-3	3.89	62.86
OpenNMT Google Big	5.03	57.56

Model Performance We used this model to generate 'reconstructed' ligand fingerprints based solely on interpretation of protein pocket fingerprints. For assessment, we translated the beam-search generated sequence into a fingerprint representation, and compared this reconstructed fingerprint to that calculated from the crystal structure (see subsection 3.6.3). We found that the model was capable of producing a significant improvement (Figure 4.33a) on the length-adjusted baseline reconstruction performance (Tc = 0.04), with 18.4% of generated fingerprints over the corresponding cutoff threshold for FDFP6 (Equivalent parameters for FDFP6 to the models described in Table 4.21 are: training MCC 0.23±0.02, training recall 0.08±0.02, testing MCC 0.41, testing recall 0.61, threshold 0.77). The model results in Tc $\geq$ 0.5 in more than two-thirds (65.7%) of cases, with a median similarity of 0.57.

4.5.2 Discussion

Given the known importance of local roughness and fine shape in determining binding site selectivity, we sought to train a model to infer complementary ligand shape representations from those of identified binding sites. As previously discussed, earlier works in this area found no straightforward relationship when the binding partners are consid-

Table 4.25: Validation set perplexity and accuracy for all combinations of parameters tested for the AIAYN Keras transformer implementation. Results are grouped by optimiser, depth and width, and, lastly number of attention heads. Best results per optimiser row and column are highlighted. Overall, Adam has the best overall performance on this task, then Nadam, and finally the Adadelta approach. For Adam, no very significant difference is observed between the configurations trialled, although beyond a certain complexity threshold, we notice a degradation in performance. Nadam shows a similar pattern. Adadelta degrades much faster in this case, with increasing parameter count. Greyed-out boxes indicate combinations for which memory requirements were exceeded, or convergence was not reached after 2000 epochs. All results shown are for FDFP6, 512-bit representation.

Optimiser	**Layers, Width**	**4**	**8**	**12**	**16**
Adam	**2 256**	2.19, 75.08	2.16, 75.48	2.18, 75.34	2.21, 74.89
	2 512	2.16, 75.44	2.18, 75.68	2.17, 75.47	2.22, 75.35
	4 256	2.16, 75.52	2.15, 75.68	2.15, 75.34	2.19, 75.34
	4 512	2.15, 76.02	**2.14, 75.97**	**2.15, 75.90**	2.19, 75.65
	6 256	**2.14, 75.76**	2.16, 75.74	2.15, 75.85	**2.18, 75.53**
	6 512	2.17, 75.66	2.19, 75.69	2.19, 75.68	2.19, 75.67
	8 256	2.17, 75.64	2.18, 75.56	2.18, 75.62	2.9, 63.03
	8 512	3.24, 61.35	2.25, 75.19	3.21, 61.78	3.12, 61.43
Nadam	**2 256**	2.19, 75.18	2.2, 75.04	2.2, 75.15	2.75, 63.68
	2 512	2.22, 75.09	2.22, 75.26	2.22, 75.23	2.76, 63.83
	4 256	2.19, 75.28	2.18, 75.38	2.18, 75.51	
	4 512	2.21, 75.23	2.19, 75.38	2.20, 75.48	
	6 256	**2.19, 75.52**	**2.17, 75.60**	**2.17, 75.7**	
	6 512	2.91, 67.49	2.6, 67.48	2.59, 67.17	
Adadelta	**2 256**	4.25, 59.39	4.52, 58.32	4.27, 59.31	
	2 512	**4.14, 59.85**	**4.12, 59.39**	**4.19, 59.81**	
	4 256	8.02, 37.37	9.00, 34.11	8.99, 34.53	
	4 512	8.04, 38.92	8.16, 37.41	8.45, 36.02	
	6 256	9.53, 29.75	10.52, 26.52	10.54, 26.24	
	6 512	9.30, 29.34	36.87, 6.19		

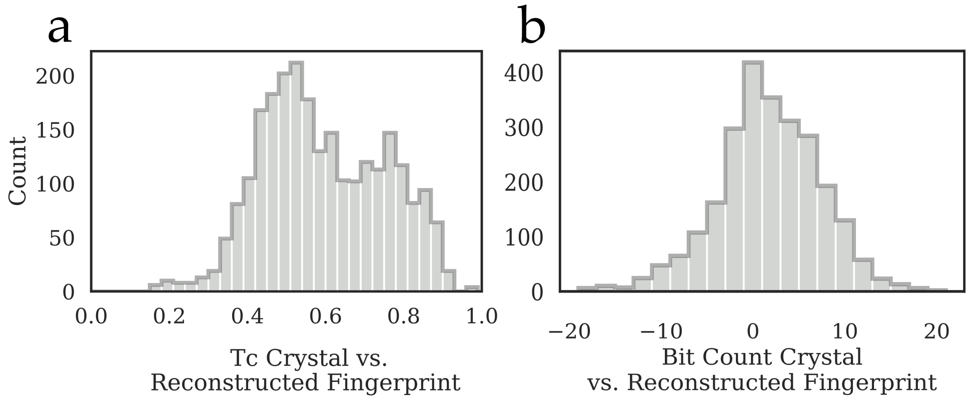


Figure 4.33: a) Pairwise similarity between crystal-extracted ligand fingerprint and that reconstructed from consideration of the protein ligand fingerprint. Median pairwise similarity is 0.57. b) Difference in the bit-count between the crystal fingerprint and the reconstructed representation. The median difference is equal to one.

ered at the whole-pocket, whole-ligand level, so we considered the issue at an atomic resolution, as with the works of Pettit [164] and Todoroff [169]. Distinguishing our approach from these methods, we use the developed shape-fingerprint methods to capture information about substructures in both binding partners, encoding these into fixed-length vectors representing the local shape properties of the ligand and associated binding pocket. These authors showed that the local roughness of a surface correlates with its ligandability, in a complex fashion. In our work, we aimed to see whether we could to some extent invert the intention of Todoroff's model; instead of finding ligandable sites on a target, we aim to translate a known ligandable site into a potential ligand, to determine the extent to which target and ligand geometries are co-determined. Finding considerable variability in binding pocket, and bound ligand, geometries (Figure 4.31) suggested that such an approach might be feasible, especially given that we had already found the ligand FDFP representation useful for 'real-world' LBVS campaigns.

To the best of our knowledge, this is the first attempt to directly translate a binding pocket shape representation into a ligand format useful for virtual screening approaches. While significant work has been done to date on the subject of shape-based complementarity, these approaches have typically been evaluating existing pairs [132, 489, 490], or driving a geometric docking approach [375], with a few exceptions [271, 491]. The latter cases each directly incorporate electrostatic information, and were both validated on the DUD dataset, and compared against other structure-based virtual screening methods. Our finding that the developed proof-of-concept model could accurately recreate ligand fingerprints, and, in a fifth of cases, sufficiently well that we would retrieve the original ligand with high-confidence of shared activity based solely on an analysis of the protein structure, is very encouraging.

Looking into the errors in more detail, we find that, more often than not, the generated fingerprint has fewer bits than the original crystal form (Figure 4.33b). One benefit of fingerprint methods is their extension of atom descriptions to local neighbourhoods. It could certainly be the case that one reason for this discrepancy is that only the ‘near-side atoms’, by which we mean those with target-facing surfaces, have shape fingerprint-bits set in the output. Whilst the overall performance of the model is surprisingly good, further optimisation of both the engineered features (fingerprints) and the model hyperparameters, and visual analysis of the cases at the extremes of the performance distribution, might help direct improvements to shift more of the distribution beyond the threshold. Additionally, as noted by Ehrt et al. [148], the choice of similarity metric used to compare binding pocket shape has a substantial impact on the quality of results obtained.

We consider there to be seven significant sources of potential error in our approach:

1. Protein-description: Given that we have no adequate system for the benchmarking of protein pocket similarity, it is plausible that our fingerprinting method fails to capture sufficient information about pocket geometry, either due to an intrinsic effect, or due to parametrisation. Free parameters, such as the size of the local environment, and the distance cut-off for atoms near the bound ligand, are sources of the latter error. Other local-shape methods for macromolecules have been proposed [168], but this field lacks a standardised benchmark such as that discussed for small molecules.

2. Environment-description: the simplification of most immediate biological relevance is our removal of solvents and co-factors. The importance of these for determining binding specificity has been discussed extensively in the literature [492–496]. We ignored these for reasons of computational expediency.

3. Target and ligand dynamics: Our approach is based on a 'snapshot' of a target state, under certain conditions, for which a model is constructed satisfying the electron density observed. We made no explicit efforts to further curate the data, or to account for intrinsic variability in the structures themselves. As noted in Table 4.23, comparing a crystallographic ligand fingerprint with those describing generated conformers results in some loss of similarity (Tc 0.84±0.04). The impact of protein structure on ligand conformation, and vice versa, is innately important for shape considerations. Analysis of molecular dynamics trajectories, as with the work of Todoroff et al. [169], allowing for the construction of a consensus representation, might be a profitable direction for further research.

4. Fingerprint generation (I): One explanation for the good performance of the transformer architecture on this task is its lack of strong position-dependence. As constructed, we simply transcribe the index positions of on-bits to a string representation. This does not take into account the intrinsic order of bit-setting, i.e., it is inherently agnostic to whether or not these bits were set in the first or last iteration of the fingerprinting algorithm, and hence what level or size of feature is being described. Additionally, we do not distinguish between iterations when considering the retrospective performance. While the transformer architecture has shown promise for chemistry-related tasks, such as reaction prediction

[497], SMILES are a more 'natural' grammar than that adopted here.

5. Fingerprint generation (II): Equally, the fingerprinting algorithm used for each method differs. As previously described, E3FP-like algorithms (protein fingerprint) generate fingerprints based on a radius in Euclidean, rather than topological space. Repeating this experiment with an E3FP-like ligand fingerprint would help to lessen this representational error.
6. Model-capacity and training: Given that we observed a clear improvement through use of the transformer architecture, it is probable that we have not identified the best-possible architecture for this problem. Models with a higher capacity to generalise, or which more directly represent the pocket, such asc a 3D-CNN [358, 498], might be more appropriate for this task.
7. Intrinsic error: As discussed previously, it is unlikely that target and ligand geometry suffice to determine their interactions entirely. Electrostatic effects and crystallographic considerations likely impose an upper limit on the achieved performance with such an approach. Shin et al. [271] noted substantial improvements in performance of their method with the addition of electrostatics. However, such an approach would render it quite difficult to assess the contribution of the geometric component.

Given that our extensive hyperparameter optimisation, and assessment of models of varying complexity, did not substantially affect performance, we feel that priority should be given to the definition of the pocket atoms, and focused analysis of the fingerprint bits set in the output. Colourisation of the original atoms with the subsequent bit indices is relatively straightforward, and would help to guide further efforts in this matter to focus efforts.

Utilisation of the crystal-derived and reconstructed fingerprints in a similar benchmarking approach to that described could be useful to determine whether Tc reduction is associated with real-world performance degradation, as would attempting this transductive process with a fingerprint derived from an apo-structure. In addition to demonstrating the feasibility of ligand fingerprint suggestion, based on pocket shape, we also see the issues associated with the direct adoption of metrics considered common in machine-learning domains for cheminformatics. Validation perplexity and accuracy for our model have more direct meaning in language translation tasks. Utilising these directly in chemoinformatics, where small alterations can interfere substantially with similarity searching, for example, is unlikely to be as useful as translating the results into a more use-case appropriate format.

Chapter 5

Conclusion

Hypothesis: 'Shape complementarity determines a considerable portion of the target-complementarity of ligands, and the ligand-complementarity of targets, in a manner substantively different to simple descriptors.'

This work aimed to explore the geometric complementarity between ligands and targets from a variety of perspectives. This has been extensively discussed in the literature as contributing to the specificity of both partners in that interaction. Existing approaches have been widely adopted in academic and industrial drug-discovery settings, illustrating the descriptive power of such methods. Previous work has made much of the capacity of these methods to facilitate scaffold-hopping, and to predict the promiscuity of a scaffold [499]. Existing implementations of this idea tend towards one of two extremes; exceptional speed and approximate description, or a rigorous definition of similarity and poor scaling characteristics.

Given previous work in the field of small- and macro-molecular comparison, based on the α-shape formalism, and recent discussion of the ideal properties of description in terms of fractal dimensionality for the latter group, we sought to investigate whether we could harness this formalism to facilitate rapid, rigorous, shape-based descriptions of large swathes of chemical space. With an eye to efficient algorithm development, but without speed as a primary focus, the developed approaches (global and local fractal dimensionality descriptors) have exceptional run-time performance, enabling shape-based comparisons at a speed typically associated with 2D methods. There is plenty of room for further improvement in the design and implementation of the approach, which provides a natural and readily-extensible framework. Our method is suitable by default for a wide range of chemical matter, from small organic molecules, to local and global descriptions of their macromolecular binding partners.

To assess the relevance of the described approaches, we adopted an external, large-scale benchmarking strategy. This furnished some interesting insights into the relationship between 2D and 3D chemical space. We observed that shape-based methods, including FDFP and GFD, behave substantially differently to their 2D cousins when challenged with different datasets. The MUV dataset, designed to be challenging for the familiar 2D approaches, proved easier for each of the shape-based approaches assessed. This relationship was inverted when considering a subset of the ChEMBL database, which represents the fruits of years of labour in drug development. Here, 3D methods typically underperformed, although FDFP was observed to retain good performance levels. Given this disparity, one reasonable interpretation of the limitations observed with such approaches is that our existing libraries of validated compounds are not 'very 3D' in nature. This is supported by analyses of the diversity of screening libraries [82, 499], and it to some extent a reflection of the easiest-to-reach parts of chemical space, which are therefore most densely covered in synthesised molecules.

To strike out of this comfort zone, out of the flatlands and into the hills, it is reassuring to have some sort of guide. In our case, we adopted a multi-faceted approach, where we took insights from the benchmarking study, and combined these with two prospective studies, to assess the applicability domain of our approach. The first of these prospective studies resulted in the identification of a novel inhibitor of TRPC4, representing a significant scaffold hop from the template, the natural product (-)-Englerin A. A small SAR study demonstrated both the power and limitations of a global shape-based approach. A second, much larger, second study aimed to assess the relationship between shape- and biological-similarity of compounds for seven targets. We saw that insights learned from a 2D benchmarking study could be utilised to enrich a 3D VS campaign, resulting in a hit rate of 42% overall beyond a cutoff threshold identified in that first study. These compounds represent a diverse sampling of chemical space, with a high proportion of unique scaffolds, previously unknown in public databases. As such, we find evidence in support of our underlying hypothesis, that shape is in some way privileged, as we can make a jump in pharmacophore and scaffold spaces without falling off an activity cliff.

Examining the shape relationship between small- and macro-molecules at a finer level of granularity, we conducted an extensive profile of a promising compound, active against the PIM1 kinase. We collated data on the specificity of kinase inhibitors, and obtained a crystal structure for our compound in complex with PIM1. Using this as the basis of a focused set, we assessed the performance of various shape-based strategies for the comparison of bound and free ligand conformations, and, intriguingly, noted some correlation between shape similarity in ligand and target spaces, although this was insufficient to adequately discriminate between known ligands of PIM1, given that they were all sited within the canonical binding pocket.

To further investigate that potential lead, we attempted a proof-of-concept study, demonstrating the 'translation' of target pockets into complementary ligands through a variety of AI approaches. We described roughly fourteen thousand protein-ligand pairs in terms of their shape, and profiled various approaches which have been employed successfully in other fields, finally settling on a relatively novel architecture. We found that the method worked sufficiently well to suggest screening-ready shape fingerprints in a quarter of cases, and in two-thirds we recovered more than a half of the fingerprint, enabling a fuzzier search. As such, this approach appears to work in principle, although future efforts are needed to clarify the underlying factors determining success for a given pocket. Recent work [357] has demonstrated the desire to incorporate considerations of shape in generative efforts, and we feel our work is a step towards that, albeit taking a different approach. As with that work, incorporation of pharmacophoric features, as with ROCS Colour and USRCAT, could add valuable information, and improve the specificity of our representation. The FDFP approach is sufficiently fast that its incorporation into generative models as a scoring metric, guiding the focused generation of a library of shape-similar compounds, would be straightforward.

Returning to our initial hypothesis, we feel that the insights gained from the benchmarking and prospective studies offer strong support for the statement above, demonstrating that shape-based methods can perform equally well, or better, on datasets constructed using traditional approaches, identifying a higher proportion of novel scaffolds in both retrospective and prospective applications than would be expected. For the moment, support for the 'ligand-complementarity of targets' is predicated on the ligand-based characterisation of target space, a traditional approach in pharmacological endeavours [500], although we have made some early efforts directly supporting this clause.

Further work to support this, and to help parametrise the protein FDFP approach, could utilise the constructed dataset of protein fingerprints to build similarity networks of protein fingerprints, and consider the enrichment in gene ontology [501, 502] terms, allowing for the direct correlation of shape and functional similarity [148]. In a similar vein, the developed FDFP approach is sufficiently fast and powerful to facilitate broader comparisons in chemical as well as biological space. A simple extension of the analysis here would be to cluster screening libraries in terms of their shape, providing a useful analysis of the diversity of 'geotypes', by analogy to chemotypes [503], present in any given set. Such an analysis could also be used to determine the overlap of these two concepts, i.e. the extent to which members of a cluster are isofunctional, and, by extension, to map between ligand and protein shape-clusters. Although we showed that a simple FDFP similarity model performed well as a basic form of target prediction, this would be a natural direction to take for further development.

Given the observation that structure is more conserved than sequence for the binding sites of protein targets [165, 504, 505], and that a ligand's shape is an adequate predictor of its promiscuity [499], an extension of our analysis with the PIM1 kinase could prove fruitful, given that we were unable to find sufficient overlap in our specificity and crystallographic datasets to adequately explore this relationship. A simple extension would be the incorporation of crystallographic data for the other kinases profiled, to assess the overlap between pocket similarity and comparable bioactivity.

Overall, considering the various studies conducted, one could say, cautiously, that the picture which emerges is one of shape providing an underlying framework, upon which pharmacophoric features act as determinants of specificity. This explains the observation that high shape-similarity is not as clear an indication of shared activity as high 2D similarity, but that the latter approach alone is most useful in areas of chemical space where we are already confident of joint activity, and has less utility for scaffold-hopping. Equally, we find some support for the idea that target shape and ligand shape are intrinsically related, and in a complementary fashion. The falsifiability of our hypothesis is uncertain, given the difficulty of separating geometry and electrostatics, but we have found little substantial evidence against it in this work, and much in its favour. As such, we feel that shape analysis has a valuable role to play in investigating the nature of ligand-target interactions, and in the identification of novel bioactive compounds based on the principle of shape-similarity. The developed approach extends the field by means of a rigorous definition of shape, which is computable in a reasonable time, and is readily extensible to the description of new chemical and biological entities, facilitating further investigation of this concept.

Bibliography

1. Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J Chem Inf Comput Sci* **43,** 374–380 (2003).
2. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des* **27,** 675–679 (2013).
3. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* **16,** 3–50 (1996).
4. 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* **52,** 2864–2875 (2012).
5. Schamberger, J., Grimm, M., Steinmeyer, A. & Hillisch, A. Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. *Drug Discov Today* **16,** 636–641 (2011).
6. Bleicher, K. H., Böhm, H.-J., Müller, K. & Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* **2,** 369–378 (2003).
7. Jacoby, E. *et al.* Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr Top Med Chem* **5,** 397–411 (2005).

8. Zhu, T. *et al.* Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis. *J Med Chem* **56,** 6560–6572 (2013).

9. Ferreira, R. S. *et al.* Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. *J Med Chem* **53,** 4891–4905 (2010).

10. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br J Pharmacol* **162,** 1239–1249 (2011).

11. Inglese, J. *et al.* Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc Natl Acad Sci USA* **103,** 11473–11478 (2006).

12. Zaragoza-Sundqvist, M., Eriksson, H., Rohman, M. & Greasley, P. J. High-quality cost-effective compound management support for HTS. *J Biomol Screen* **14,** 509–514 (2009).

13. Clark, M. A. *et al.* Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat Chem Biol* **5,** 647–654 (2009).

14. Paricharak, S., IJzerman, A. P., Bender, A. & Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data. *ACS Chem Biol* **11,** 1255–1264 (2016).

15. Williams, K. *et al.* Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J R Soc Interface* **12,** 20141289 (2015).

16. Volochnyuk, D. M. *et al.* Evolution of commercially available compounds for HTS. *Drug Discov Today* **24,** 390–402 (2019).

17. Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discov* **14,** 475–486 (2015).

18. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov* **11,** 191–200 (2012).

19. Kogej, T. *et al.* Big pharma screening collections: more of the same or unique libraries? The AstraZeneca-Bayer Pharma AG case. *Drug Discov Today* **18,** 1014–1024 (2013).

20. Engels, M. F. M. *et al.* A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J Chem Inf Model* **46,** 2651–2660 (2006).

21. Weber, L. High-diversity combinatorial libraries. *Curr Opin Chem Biol* **4,** 295–302 (2000).

22. Feher, M. & Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci* **43,** 218–227 (2003).

23. Lipkin, M. J., Stevens, A. P., Livingstone, D. J. & Harris, C. J. How large does a compound screening collection need to be? *Comb Chem High Throughput Screen* **11,** 482–493 (2008).

24. Harris, C. J., Hill, R. D., Sheppard, D. W., Slater, M. J. & Stouten, P. F. W. The design and application of target-focused compound libraries. *Comb Chem High Throughput Screen* **14,** 521–531 (2011).

25. Lipkus, A. H. *et al.* Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J Org Chem* **73,** 4443–4451 (2008).

26. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **39,** 2887–2893 (1996).

27. Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287,** 1964–1969 (2000).

28. Schreiber, S. L. Molecular diversity by design. *Nature* **457,** 153–154 (2009).

29. Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat Chem* **8,** 531–541 (2016).

30. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* **4,** 206–220 (2005).

31. Galloway, W. R. & Spring, D. R. Is synthesis the main hurdle for the generation of diversity in compound libraries for screening? *Expert Opin Drug Discov* **4,** 467–472 (2009).

32. Renner, S. *et al.* Recent trends and observations in the design of high-quality screening collections. *Future Med Chem* **3,** 751–766 (2011).

33. Hu, Y. & Bajorath, J. Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database. *J Chem Inf Model* **52,** 1806–1811 (2012).

34. Stumpfe, D. & Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J Med Chem* **55,** 2932–2942 (2012).

35. Schneider, P. & Schneider, G. De novo design at the edge of chaos. *J Med Chem* **59,** 4077–4086 (2016).

36. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov* **3,** 353–359 (2004).

37. Peters, J.-U. Polypharmacology - foe or friend? *J Med Chem* **56,** 8955–8971 (2013).

38. Selvam, B., Porter, S. L. & Tikhonova, I. G. Addressing selective polypharmacology of antipsychotic drugs targeting the bioaminergic receptors through receptor dynamic conformational ensembles. *J Chem Inf Model* **53,** 1761–1774 (2013).

39. Anighoro, A., Bajorath, J. & Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery. *J Med Chem* **57,** 7874–7887 (2014).

40. Hansch, C., Maloney, P. P., Fujita, T. & Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194,** 178–180 (1962).

41. Sheridan, R. P. & Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov Today* **7,** 903–911 (2002).

42. MDL Information Systems, I. *MACCS Keys* 14600 Catalina Street, San Leandro, CA 94577., 1996.

43. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50,** 742–754 (2010).

44. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew Chem Int Ed Engl* **38,** 2894–2896 (1999).

45. Reutlinger, M. *et al.* Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol Inform* **32,** 133–138 (2013).

46. Wermuth, C. G., Ganellin, C. R., Lindberg, P. & Mitscher, L. A. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **70,** 1129–1143 (1998).

47. Vainio, M. J., Puranen, J. S. & Johnson, M. S. ShaEP: molecular overlay based on shape and electrostatic potential. *J Chem Inf Model* **49,** 492–502 (2009).

48. Rush, 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* **48,** 1489–1495 (2005).

49. Hopfinger, A. J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J Am Chem Soc* **102,** 7196–7206 (1980).

50. Tosco, P., Balle, T. & Shiri, F. Open3DALIGN: an open-source software aimed at unsupervised ligand alignment. *J Comput Aided Mol Des* **25,** 777–783 (2011).

51. Pavanï, R. & Ranghino, G. A method to compute the volume of a molecule. *Comput Chem* **6,** 133–135 (1982).

52. Connolly, M. L. Computation of molecular volume. *J Am Chem Soc* **107,** 1118–1124 (1985).

53. Masek, B. B., Merchant, A. & Matthew, J. B. Molecular shape comparison of angiotensin II receptor antagonists. *J Med Chem* **36,** 1230–1238 (1993).

54. Masek, B. B., Merchant, A. & Matthew, J. B. Molecular skins: a new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins* **17,** 193–202 (1993).

55. Grant, J. A. & Pickup, B. T. A Gaussian Description of Molecular Shape. *J Phys Chem* **99,** 3503–3510 (1995).

56. 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* **17,** 1653–1666 (1996).

57. Carbó, R., Leyda, L. & Arnau, M. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int J Quantum Chem* **17,** 1185–1189 (1980).

58. Good, A. C., Hodgkin, E. E. & Richards, W. G. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J Chem Inf Model* **32,** 188–191 (1992).

59. Good, A. C. & Richards, W. G. Rapid evaluation of shape similarity using Gaussian functions. *J Chem Inf Model* **33,** 112–116 (1993).

60. Boys, S. Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system. *Proc. R. Soc. Lond. A* **200,** 542–554 (1950).

61. Nauchitel, V. V. & Somorjai, R. L. Gaussian neighborhood: a new measure of accessibility for residues of protein molecules. *Proteins* **15,** 50–61 (1993).

62. Yan, X. *et al.* Enhancing molecular shape comparison by weighted Gaussian functions. *J Chem Inf Model* **53,** 1967–1978 (2013).

63. Yan, X., Li, J., Gu, Q. & Xu, J. gWEGA: GPU-accelerated WEGA for molecular superposition and shape comparison. *J Comput Chem* **35,** 1122–1130 (2014).

64. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* **50,** 74–82 (2007).

65. Kearnes, S. & Pande, V. ROCS-derived features for virtual screening. *J Comput Aided Mol Des* **30,** 609–617 (2016).

66. Sato, T. *et al.* Application of support vector machine to three-dimensional shape-based virtual screening using comprehensive three-dimensional molecular shape overlay with known inhibitors. *J Chem Inf Model* **52,** 1015–1026 (2012).

67. Haque, I. S. & Pande, V. S. PAPER–accelerating parallel evaluations of ROCS. *J Comput Chem* **31,** 117–132 (2010).

68. Vaz de Lima, L. A. C. & Nascimento, A. S. MolShaCS: a free and open source tool for ligand similarity identification based on Gaussian descriptors. *Eur J Med Chem* **59,** 296–303 (2013).

69. Liu, X., Jiang, H. & Li, H. SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J Chem Inf Model* **51,** 2372–2385 (2011).

70. Sastry, G. M., Dixon, S. L. & Sherman, W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J Chem Inf Model* **51,** 2455–2466 (2011).

71. Hamza, A., Wei, N.-N. & Zhan, C.-G. Ligand-based virtual screening approach using a new scoring function. *J Chem Inf Model* **52,** 963–974 (2012).

72. Hamza, A., Wei, N.-N., Hao, C., Xiu, Z. & Zhan, C.-G. A novel and efficient ligand-based virtual screening approach using the HWZ scoring function and an enhanced shape-density model. *J Biomol Struct Dyn* **31,** 1236–1250 (2013).

73. Proschak, E., Rupp, M., Derksen, S. & Schneider, G. Shapelets: possibilities and limitations of shape-based virtual screening. *J Comput Chem* **29,** 108–114 (2008).

74. Proschak, E. *et al.* From molecular shape to potent bioactive agents I: bioisosteric replacement of molecular fragments. *ChemMedChem* **4,** 41–44 (2009).

75. Proschak, E. *et al.* From molecular shape to potent bioactive agents II: fragment-based de novo design. *ChemMedChem* **4,** 45–48 (2009).

76. Gramada, A. & Bourne, P. E. Multipolar representation of protein structure. *BMC Bioinformatics* **7,** 242 (2006).

77. Konarev, P. V., Petoukhov, M. V. & Svergun, D. I. Rapid automated superposition of shapes and macromolecular models using spherical harmonics. *J Appl Crystallogr* **49,** 953–960 (2016).

78. Morris, R. J., Najmanovich, R. J., Kahraman, A. & Thornton, J. M. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **21,** 2347–2355 (2005).

79. Karaboga, A. S., Petronin, F., Marchetti, G., Souchet, M. & Maigret, B. Benchmarking of HPCC: A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments. *J Mol Graph Model* **41,** 20–30 (2013).

80. Mavridis, L., Hudson, B. D. & Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic surface representations. *J Chem Inf Model* **47,** 1787–1796 (2007).

81. Wang, Q. *et al.* Spherical harmonics coefficients for ligand-based virtual screening of cyclooxygenase inhibitors. *PLoS ONE* **6,** e21554 (2011).

82. Sauer, W. H. B. & Schwarz, M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J Chem Inf Comput Sci* **43,** 987–1003 (2003).

83. Nilakantan, R., Bauman, N. & Venkataraghavan, R. New method for rapid characterization of molecular shapes: applications in drug design. *J Chem Inf Model* **33,** 79–85 (1993).

84. Bemis, G. W. & Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J Comput Aided Mol Des* **6,** 607–628 (1992).

85. Good, A. C., Ewing, T. J., Gschwend, D. A. & Kuntz, I. D. New molecular shape descriptors: application in database screening. *J Comput Aided Mol Des* **9,** 1–12 (1995).

86. Gionis, A., Indyk, P. & Motwani, R. Similarity search in high dimensions via hashing. *Vldb* **99,** 518 (1999).

87. Ballester, P. J. & Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem* **28,** 1711–1723 (2007).

88. Ballester, P. J., Westwood, I., Laurieri, N., Sim, E. & Richards, W. G. Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. *J R Soc Interface* **7,** 335–342 (2010).

89. Ballester, P. J. *US Patent* 8244483 (2012).

90. Kumar, A. & Zhang, K. Y. J. Advances in the development of shape similarity methods and their application in drug discovery. *Front Chem* **6,** 315 (2018).

91. Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. J. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Res* **44,** W436–41 (2016).

92. Sterling, T. & Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J Chem Inf Model* **55,** 2324–2337 (2015).

93. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **52,** 1757–1768 (2012).

94. Patil, S. P., Ballester, P. J. & Kerezsi, C. R. Prospective virtual screening for novel p53-MDM2 inhibitors using ultrafast shape recognition. *J Comput Aided Mol Des* **28,** 89–97 (2014).

95. Teo, C. *et al.* Ligand-based virtual screening for the discovery of inhibitors for protein arginine deiminase type 4 (PAD4). *Metabolomics* **3,** 4 (2013).

96. Cannon, E. O., Nigsch, F. & Mitchell, J. B. O. A novel hybrid ultrafast shape descriptor method for use in virtual screening. *Chem Cent J* **2,** 3 (2008).

97. Schreyer, A. & Blundell, T. CREDO: a protein-ligand interaction database for drug discovery. *Chem Biol Drug Des* **73,** 157–167 (2009).

98. Schreyer, A. M. & Blundell, T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J Cheminform* **4,** 27 (2012).

99. Armstrong, M. S. *et al.* ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J Comput Aided Mol Des* **24,** 789–801 (2010).

100. Armstrong, M. S., Finn, P. W., Morris, G. M. & Richards, W. G. Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J Comput Aided Mol Des* **25,** 785–790 (2011).

101. Berenger, F., Voet, A., Lee, X. Y. & Zhang, K. Y. A rotation-translation invariant molecular descriptor of partial charges and its use in ligand-based virtual screening. *J Cheminform* **6,** 23 (2014).

102. Wagener, M., Sadowski, J. & Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks. *J Am Chem Soc* **117,** 7769–7775 (1995).

103. Broto, P., Moreau, G. & Vandycke, C. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *European Journal of Medicinal Chemistry* **19,** 71–78 (1984).

104. Todeschini, R. & Gramatica, P. SD-modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act.Relat.* **16,** 113–119 (1997).

105. Todeschini, R. & Gramatica, P. The whim theory: new 3D molecular descriptors for qsar in environmental modelling. *SAR QSAR Environ Res* **7,** 89–115 (1997).

106. Axen, S. D. *et al.* A Simple Representation of Three-Dimensional Molecular Structure. *J Med Chem* **60,** 7393–7409 (2017).

107. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* **110,** 5959–5967 (1988).

108. Klebe, G., Abraham, U. & Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* **37,** 4130–4146 (1994).

109. Silverman, B. D. & Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J Med Chem* **39,** 2129–2140 (1996).

110. Pastor, M., Cruciani, G., McLay, I., Pickett, S. & Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* **43,** 3233–3243 (2000).

111. Damale, M., Harke, S., Kalam Khan, F., Shinde, D. & Sangshetti, J. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. *MRMC* **14,** 35–55 (2014).

112. Goodford, P. in *Molecular interaction fields: applications in drug discovery and ADME prediction* (ed Cruciani, G.) 1–25 (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, 2005).

113. Verma, J., Khedkar, V. M. & Coutinho, E. C. 3D-QSAR in drug design–a review. *Curr Top Med Chem* **10,** 95–115 (2010).

114. Kubinyi, H., Folkers, G. & Martin, Y. *3D QSAR in drug design: recent advances* (Springer Science & Business Media, 2006).

115. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55,** 379–400 (1971).

116. Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology* **79,** 351–371 (1973).

117. Connolly, M. L. Analytical molecular surface calculation. *J Appl Crystallogr* **16,** 548–558 (1983).

118. Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221,** 709–713 (1983).

119. Goldman, B. B. & Wipke, W. T. Quadratic shape descriptors. 1. rapid superposition of dissimilar molecules using geometrically invariant surface descriptors. *J Chem Inf Comput Sci* **40,** 644–658 (2000).

120. Cosgrove, D. A., Bayada, D. M. & Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J Comput Aided Mol Des* **14,** 573–591 (2000).

121. Hofbauer, C., Lohninger, H. & Aszódi, A. SURFCOMP: a novel graph-based approach to molecular surface comparison. *J Chem Inf Comput Sci* **44,** 837–847 (2004).

122. Zauhar, R. J., Moyna, G., Tian, L., Li, Z. & Welsh, W. J. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J Med Chem* **46,** 5674–5690 (2003).

123. Meek, P. J. *et al.* Shape Signatures: speeding up computer aided drug discovery. *Drug Discov Today* **11,** 895–904 (2006).

124. Werner, M. M., Li, Z. & Zauhar, R. J. Computer-aided identification of novel 3,5-substituted rhodanine derivatives with activity against Staphylococcus aureus DNA gyrase. *Bioorg Med Chem* **22,** 2176–2187 (2014).

125. Zauhar, R. J., Gianti, E. & Welsh, W. J. Fragment-based Shape Signatures: a new tool for virtual screening and drug discovery. *J Comput Aided Mol Des* **27,** 1009–1036 (2013).

126. Liu, Y.-S., Fang, Y. & Ramani, K. IDSS: deformation invariant signatures for molecular shape comparison. *BMC Bioinformatics* **10,** 157 (2009).

127. Edelsbrunner, H., Kirkpatrick, D. & Seidel, R. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory* **29,** 551–559 (1983).

128. Edelsbrunner, H. & Mücke, E. P. Three-dimensional alpha shapes. *ACM Trans Graph* **13,** 43–72 (1994).

129. Wilson, J. A., Bender, A., Kaya, T. & Clemons, P. A. Alpha shapes applied to molecular shape characterization exhibit novel properties compared to established shape descriptors. *J Chem Inf Model* **49,** 2231–2241 (2009).

130. Edelsbrunner, H. & Shah, N. Incremental topological flipping works for regular triangulations. *Algorithmica* **15,** 223 (1996).

131. Edelsbrunner, H. & Mücke, E. P. *Three-dimensional alpha shapes* in *Proceedings of the 1992 workshop on Volume visualization - VVS '92* (ACM Press, New York, New York, USA, 1992), 75–82.

132. Kahraman, A., Morris, R. J., Laskowski, R. A. & Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J Mol Biol* **368,** 283–301 (2007).

133. Pérot, S., Sperandio, O., Miteva, M. A., Camproux, A.-C. & Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov Today* **15,** 656–667 (2010).

134. Nayal, M. & Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **63,** 892–906 (2006).

135. Laurie, A. T. R. & Jackson, R. M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr Protein Pept Sci* **7,** 395–406 (2006).

136. Zheng, X., Gan, L., Wang, E. & Wang, J. Pocket-based drug design: exploring pocket space. *AAPS J* **15,** 228–241 (2013).

137. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci* **5,** 2438–2452 (1996).

138. Rosen, M., Lin, S. L., Wolfson, H. & Nussinov, R. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng* **11,** 263–277 (1998).

139. Bergner, A. & Günther, J. in *Chemogenomics in drug discovery: A medicinal chemistry perspective* (eds Kubinyi, H. & Müller, G.) 97–135 (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, 2004).

140. Gold, N. D. & Jackson, R. M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J Mol Biol* **355,** 1112–1124 (2006).

141. An, J., Totrov, M. & Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* **4,** 752–761 (2005).

142. Beinglass, A. & Wolfson, H. *Articulated object recognition, or: how to generalize the generalized Hough transform* in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Comput. Sco. Press, 1991), 461–466.

143. Sandak, B., Nussinov, R. & Wolfson, H. J. An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. *Comput Appl Biosci* **11,** 87–99 (1995).

144. Nussinov, R. & Wolfson, H. J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA* **88,** 10495–10499 (1991).

145. Xie, L., Xie, L. & Bourne, P. E. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **25,** i305–12 (2009).

146. Kellenberger, E., Schalon, C. & Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *CAD* **4,** 209–220 (2008).

147. Chaudhari, R., Tan, Z. & Zhang, S. in *Comprehensive medicinal chemistry III* 259–275 (Elsevier, 2017).

148. Ehrt, C., Brinkjost, T. & Koch, O. Binding site characterization - similarity, promiscuity, and druggability. *Medchemcomm* **10,** 1145–1159 (2019).

149. Yeturu, K. & Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics* **9,** 543 (2008).

150. Schmitt, S., Kuhn, D. & Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* **323,** 387–406 (2002).

151. Bron, C. & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun ACM* **16,** 575–577 (1973).

152. Chartier, M. & Najmanovich, R. Detection of binding site molecular interaction field similarities. *J Chem Inf Model* **55,** 1600–1615 (2015).

153. Chartier, M., Adriansen, E. & Najmanovich, R. IsoMIF Finder: online detection of binding site molecular interaction field similarities. *Bioinformatics* **32,** 621–623 (2016).

154. Kinoshita, K., Furui, J. & Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J Struct Funct Genomics* **2,** 9–22 (2002).

155. Jambon, M., Imberty, A., Deléage, G. & Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **52,** 137–145 (2003).

156. Jambon, M. *et al.* The SuMo server: 3D search for protein functional sites. *Bioinformatics* **21,** 3929–3930 (2005).

157. Weisel, M., Proschak, E. & Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Cent J* **1,** 7 (2007).

158. Weisel, M., Proschak, E., Kriegl, J. M. & Schneider, G. Form follows function: shape analysis of protein cavities for receptor-based drug design. *Proteomics* **9,** 451–459 (2009).

159. Novotni, M. & Klein, R. *3D zernike descriptors for content based shape retrieval* in *Proceedings of the eighth ACM symposium on Solid modeling and applications - SM '03* (ACM Press, New York, New York, USA, 2003), 216.

160. Kihara, D., Sael, L., Chikhi, R. & Esquivel-Rodriguez, J. Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. *Curr Protein Pept Sci* **12,** 520–530 (2011).

161. Edelsbrunner, H., Facello, M. & Liang, J. On the definition and the construction of pockets in macromolecules. *Pac Symp Biocomput,* 272–287 (1996).

162. Liang, J., Edelsbrunner, H. & Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* **7,** 1884–1897 (1998).

163. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins* **33,** 1–17 (1998).

164. Pettit, F. K. & Bowie, J. U. Protein surface roughness and small molecular binding sites. *J Mol Biol* **285,** 1377–1382 (1999).

165. Binkowski, T. A. & Joachimiak, A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC Struct Biol* **8,** 45 (2008).

166. Chen, X. *et al.* Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinformatics* **17,** 696–712 (2016).

167. Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F. & Rarey, M. Combining global and local measures for structure-based druggability predictions. *J Chem Inf Model* **52,** 360–372 (2012).

168. Albou, L.-P., Schwarz, B., Poch, O., Wurtz, J. M. & Moras, D. Defining and characterizing protein surface using alpha shapes. *Proteins* **76,** 1–12 (2009).

169. Todoroff, N. *et al.* Fractal dimensions of macromolecular structures. *Mol Inform* **33,** 588–596 (2014).

170. Kaye, B. Specification of the ruggedness and/or texture of a fine particle profile by its fractal dimension. *Powder Technol* **21,** 1–16 (1978).

171. Normant, F. & Tricot, C. Method for evaluating the fractal dimension of curves using convex hulls. *Phys Rev, A* **43,** 6518–6525 (1991).

172. Yixin, Z., Yueshan, X. & Fayao, L. *IFS fractal morphing based on coarse convex-hull* in *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference* (IEEE, 2011), 225–228.

173. Mandelbrot, B. How long is the coast of britain? Statistical self-similarity and fractional dimension. *Science* **156,** 636–638 (1967).

174. Richardson, L. The problem of contiguity: An appendix to Statistic of Deadly Quarrels. *General systems : yearbook of the Society for the Advancement of General Systems Theory* **61,** 139–187 (1961).

175. Sreenivasan, K. R. & Meneveau, C. The fractal facets of turbulence. *J Fluid Mech* **173,** 357–386 (1986).

176. Hurst, D. & Vassilicos, J. C. Scalings and decay of fractal-generated turbulence. *Phys. Fluids* **19,** 035103 (2007).

177. Esgiar, A. N., Naguib, R. N. G., Sharif, B. S., Bennett, M. K. & Murray, A. Fractal analysis in the detection of colonic cancer images. *IEEE Trans Inf Technol Biomed* **6,** 54–58 (2002).

178. Li, H., Giger, M. L., Olopade, O. I. & Lan, L. Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment. *Acad Radiol* **14,** 513–521 (2007).

179. Sebastián, M. V. & Navascués, M. A. A relation between fractal dimension and Fourier transform – electroencephalographic study using spectral and fractal parameters. *Int J Comput Math* **85,** 657–665 (2008).

180. Goldberger, A. L. *et al.* Fractal dynamics in physiology: alterations with disease and aging. *Proc Natl Acad Sci USA* **99 Suppl 1,** 2466–2472 (2002).

181. Korolj, A., Wu, H.-T. & Radisic, M. A healthy dose of chaos: Using fractal frameworks for engineering higher-fidelity biomedical systems. *Biomaterials* **219,** 119363 (2019).

182. Werner, D. & Ganguly, S. An overview of fractal antenna engineering research. *IEEE Antennas Propag. Mag.* **45,** 38–57 (2003).

183. Puente, C., Romeu, J., Pous, R., Garcia, X. & Benitez, F. Fractal multiband antenna based on the Sierpinski gasket. *Electron Lett* **32,** 1 (1996).

184. Shang, J. *et al.* Assembling molecular Sierpiński triangle fractals. *Nat Chem* **7,** 389–393 (2015).

185. Hu, J.-Y. *et al.* Symmetrically backfolded molecules emulating the self-similar features of a Sierpinski triangle. *Org Biomol Chem* (2019).

186. Mo, Y., Chen, T., Dai, J., Wu, K. & Wang, D. On-Surface Synthesis of Highly Ordered Covalent Sierpiński Triangle Fractals. *J Am Chem Soc* (2019).

187. Von Korff, M. & Sander, T. Molecular complexity calculated by fractal dimension. *Sci Rep* **9,** 967 (2019).

188. Reuveni, S., Klafter, J. & Granek, R. Dynamic structure factor of vibrating fractals: proteins as a case study. *Phys Rev E Stat Nonlin Soft Matter Phys* **85,** 011906 (2012).

189. Lewis, M. & Rees, D. C. Fractal surfaces of proteins. *Science* **230,** 1163–1165 (1985).

190. Kuhn, L. A. *et al.* The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J Mol Biol* **228,** 13–22 (1992).

191. Tejera, E., Machado, A., Rebelo, I. & Nieto-Villar, J. Fractal protein structure revisited: Topological, kinetic and thermodynamic relationships. *Physica A: Statistical Mechanics and its Applications* **388,** 4600–4608 (2009).

192. Chen, S. & Teixeira, J. Structure and fractal dimension of protein-detergent complexes. *Phys Rev Lett* **57,** 2583–2586 (1986).

193. Norel, R., Lin, S. L., Wolfson, H. J. & Nussinov, R. Shape complementarity at protein-protein interfaces. *Biopolymers* **34,** 933–940 (1994).

194. Peitgen, H.-O., Jürgens, H. & Saupe, D. in *Chaos and Fractals* 63–134 (Springer New York, New York, NY, 1992).

195. Dieudonné, J. *A History of Algebraic and Differential Topology, 1900 - 1960* (Birkhäuser Boston, Boston, 2009).

196. Cannon, J. W. & Mandelbrot, B. B. The fractal geometry of nature. *The American Mathematical Monthly* **91,** 594 (1984).

197. Falconer, K. *Fractal geometry: mathematical foundations and applications* (John Wiley & Sons, Ltd, Chichester, UK, 2003).

198. Li, J., Du, Q. & Sun, C. An improved box-counting method for image fractal dimension estimation. *Pattern Recognit* **42,** 2460–2469 (2009).

199. Greenside, H. S., Wolf, A., Swift, J. & Pignataro, T. Impracticality of a box-counting algorithm for calculating the dimensionality of strange attractors. *Phys. Rev. A* **25,** 3453–3456 (1982).

200. Grassberger, P. & Procaccia, I. Measuring the strangeness of strange attractors. *Physica D: Nonlinear phenomena* **9,** 189–208 (1983).

201. Grassberger, P. & Procaccia, I. Characterization of strange attractors. *Phys Rev Lett* **50,** 346–349 (1983).

202. Willett, P., Barnard, J. M. & Downs, G. M. Chemical Similarity Searching. *J Chem Inf Comput Sci* **38,** 983–996 (1998).

203. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7,** 20 (2015).

204. Baldi, P. & Nasr, R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J Chem Inf Model* **50,** 1205–1222 (2010).

205. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. & Weinberger, L. E. Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors. *J Med Chem* **39,** 3049–3059 (1996).

206. Delaney, J. S. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol Divers* **1,** 217–222 (1996).

207. Martin, Y. C., Kofron, J. L. & Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J Med Chem* **45,** 4350–4358 (2002).

208. Vogt, M., Wassermann, A. M. & Bajorath, J. Application of information-theoretic concepts in chemoinformatics. *Information* **1,** 60–73 (2010).

209. Vogt, M. & Bajorath, J. Introduction of the conditional correlated Bernoulli model of similarity value distributions and its application to the prospective prediction of fingerprint search performance. *J Chem Inf Model* **51,** 2496–2506 (2011).

210. Brown, N. & Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini Rev Med Chem* **6,** 1217–1229 (2006).

211. Schneider, G., Schneider, P. & Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb Sci* **25,** 1162–1171 (2006).

212. Nicholls, A. What do we know and when do we know it? *J Comput Aided Mol Des* **22,** 239–255 (2008).

213. Scior, T. *et al.* Recognizing pitfalls in virtual screening: a critical review. *J Chem Inf Model* **52,** 867–881 (2012).

214. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminform* **5,** 26 (2013).

215. Mason, S. J. & Graham, N. E. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q.J Royal Met. Soc.* **128,** 2145–2166 (2002).

216. Truchon, J.-F. & Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the 'early recognition' problem. *J Chem Inf Model* **47,** 488–508 (2007).

217. Sheridan, R. P., Singh, S. B., Fluder, E. M. & Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J Chem Inf Comput Sci* **41,** 1395–1406 (2001).

218. Böhm, H.-J., Flohr, A. & Stahl, M. Scaffold hopping. *Drug Discov Today Technol* **1,** 217–224 (2004).

219. McGaughey, G. B. *et al.* Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* **47,** 1504–1519 (2007).

220. Raymond, J. W. & Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des* **16,** 521–533 (2002).

221. Barker, E. J. *et al.* Scaffold hopping using clique detection applied to reduced graphs. *J Chem Inf Model* **46,** 503–511 (2006).

222. Cao, Y., Jiang, T. & Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **24,** i366–74 (2008).

223. Xu, Y. & Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J Chem Inf Comput Sci* **41,** 181–185 (2001).

224. Kirchmair, J., Markt, P., Distinto, S., Wolber, G. & Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes? *J Comput Aided Mol Des* **22,** 213–228 (2008).

225. Hu, Y., Stumpfe, D. & Bajorath, J. Computational exploration of molecular scaffolds in medicinal chemistry. *J Med Chem* **59,** 4062–4076 (2016).

226. Bender, A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin Drug Discov* **5,** 1141–1151 (2010).

227. Lagarde, N., Zagury, J.-F. & Montes, M. Benchmarking data sets for the evaluation of virtual ligand screening methods: review and perspectives. *J Chem Inf Model* **55,** 1297–1307 (2015).

228. *MDDR - MDL Drug Data Report - BIOVIA*

229. Willett, P. Similarity-based approaches to virtual screening. *Biochem Soc Trans* **31,** 603–606 (2003).

230. Hert, J. *et al.* Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org Biomol Chem* **2,** 3256–3266 (2004).

231. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* **11,** 1046–1053 (2006).

232. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J Med Chem* **49,** 6789–6801 (2006).

233. 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* **45,** 1369–1375 (2005).

234. Rohrer, S. G. & Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J Chem Inf Model* **48,** 704–718 (2008).

235. Rohrer, S. G. & Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model* **49,** 169–184 (2009).

236. Irwin, J. J. Community benchmarks for virtual screening. *J Comput Aided Mol Des* **22,** 193–199 (2008).

237. Hawkins, P. C. D., Warren, G. L., Skillman, A. G. & Nicholls, A. How to do an evaluation: pitfalls and traps. *J Comput Aided Mol Des* **22,** 179–190 (2008).

238. Good, A. C. & Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J Comput Aided Mol Des* **22,** 169–178 (2008).

239. 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* **55,** 6582–6594 (2012).

240. 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* **50,** 2079–2093 (2010).

241. Wang, Y. *et al.* PubChem's BioAssay Database. *Nucleic Acids Res* **40,** D400–12 (2012).

242. Ripphausen, P., Wassermann, A. M. & Bajorath, J. REPROVIS-DB: a benchmark system for ligand-based virtual screening derived from reproducible prospective applications. *J Chem Inf Model* **51,** 2467–2473 (2011).

243. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40,** D1100–7 (2012).

244. Heikamp, K. & Bajorath, J. Large-scale similarity search profiling of ChEMBL compound data sets. *J Chem Inf Model* **51,** 1831–1839 (2011).

245. Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J Med Chem* **52,** 6752–6756 (2009).

246. Breinbauer, R., Vetter, I. R. & Waldmann, H. From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. *Angew Chem Int Ed Engl* **41,** 2879–2890 (2002).

247. Koch, M. A., Breinbauer, R. & Waldmann, H. Protein structure similarity as guiding principle for combinatorial library design. *Biol Chem* **384,** 1265–1272 (2003).

248. Burke, M. D., Berger, E. M. & Schreiber, S. L. Generating diverse skeletons of small molecules combinatorially. *Science* **302,** 613–618 (2003).

249. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432,** 855–861 (2004).

250. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr Opin Struct Biol* **16,** 127–136 (2006).

251. Pérez-Nueno, V. I. & Ritchie, D. W. Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening. *J Chem Inf Model* **51,** 1233–1248 (2011).

252. Renner, S., Schwab, C. H., Gasteiger, J. & Schneider, G. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J Chem Inf Model* **46,** 2324–2332 (2006).

253. Schnecke, V. & Boström, J. Computational chemistry-driven decision making in lead generation. *Drug Discov Today* **11,** 43–50 (2006).

254. Kortagere, S., Krasowski, M. D. & Ekins, S. The importance of discerning shape in molecular pharmacology. *Trends Pharmacol Sci* **30,** 138–147 (2009).

255. Hoeger, B., Diether, M., Ballester, P. J. & Köhn, M. Biochemical evaluation of virtual screening methods reveals a cell-active inhibitor of the cancer-promoting phosphatases of regenerating liver. *Eur J Med Chem* **88,** 89–100 (2014).

256. Boström, J., Grant, J. A., Fjellström, O., Thelin, A. & Gustafsson, D. Potent fibrinolysis inhibitor discovered by shape and electrostatic complementarity to the drug tranexamic acid. *J Med Chem* **56,** 3273–3280 (2013).

257. Temml, V., Voss, C. V., Dirsch, V. M. & Schuster, D. Discovery of New Liver X Receptor Agonists by Pharmacophore Modeling and Shape-Based Virtual Screening. *J Chem Inf Model* **54,** 367–371 (2014).

258. Bassetto, M. *et al.* Shape-based virtual screening, synthesis and evaluation of novel pyrrolone derivatives as antiviral agents against HCV. *Bioorg Med Chem Lett* **27,** 936–940 (2017).

259. Ripphausen, P., Nisius, B., Peltason, L. & Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J Med Chem* **53,** 8461–8467 (2010).

260. Moffat, K., Gillet, V. J., Whittle, M., Bravi, G. & Leach, A. R. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J Chem Inf Model* **48,** 719–729 (2008).

261. Putta, S. & Beroza, P. Shapes of things: computer modeling of molecular shape in drug discovery. *Curr Top Med Chem* **7,** 1514–1524 (2007).

262. Giganti, D. *et al.* Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J Chem Inf Model* **50,** 992–1004 (2010).

263. Kirchmair, J. *et al.* How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J Chem Inf Model* **49,** 678–692 (2009).

264. Ginn, C. M., Willett, P. & Bradshaw, J. in *Virtual screening: an alternative or complement to high throughput screening?* (ed Klebe, G.) 1–16 (Kluwer Academic Publishers, Dordrecht, 2002).

265. Willett, P. Combination of similarity rankings using data fusion. *J Chem Inf Model* **53,** 1–10 (2013).

266. Gfeller, D., Michielin, O. & Zoete, V. Shaping the interaction landscape of bioactive molecules. *Bioinformatics* **29,** 3073–3079 (2013).

267. Gfeller, D. *et al.* SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res* **42,** W32–8 (2014).

268. AbdulHameed, M. D. M. *et al.* Exploring polypharmacology using a ROCS-based target fishing approach. *J Chem Inf Model* **52,** 492–505 (2012).

269. Mori, M. *et al.* Hit recycling: discovery of a potent carbonic anhydrase inhibitor by in silico target fishing. *ACS Chem Biol* **10,** 1964–1969 (2015).

270. Gao, Y. *et al.* A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery. *Biochem J* **451,** 313–328 (2013).

271. Shin, W.-H., Christoffer, C. W., Wang, J. & Kihara, D. PL-PatchSurfer2: Improved Local Surface Matching-Based Virtual Screening Method That Is Tolerant to Target and Ligand Structure Variation. *J Chem Inf Model* **56,** 1676–1691 (2016).

272. Wasserman, P. Neural computing: theory and practice. *Neural computing: theory and practice* (1989).

273. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536 (1986).

274. Rumelhart, D. E., Widrow, B. & Lehr, M. A. The basic ideas in neural networks. *Commun ACM* **37,** 87–92 (1994).

275. Hinton, G. E. How neural networks learn from experience. *Sci Am* **267,** 144–151 (1992).

276. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

277. Werbos, P. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting* (John Wiley & Sons, 1994).

278. Dreyfus, S. E. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. *Journal of Guidance, Control, and Dynamics* **13,** 926–928 (1990).

279. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

280. Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A. & Bengio, Y. Maxout networks. *arXiv preprint arXiv:1302.4389* (2013).

281. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15,** 1929–1958 (2014).

282. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), 770–778.

283. Meissner, M., Schmuker, M. & Schneider, G. Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training. *BMC Bioinformatics* **7,** 125 (2006).

284. Hochreiter, S., Bengio, Y., Frasconi, P. & Schmidhuber, J. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. (2001).

285. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9,** 1735–1780 (1997).

286. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

287. Grefenstette, E., Hermann, K., Suleyman, M. & Blunsom, P. Learning to transduce with unbounded memory. *Advances in neural information processing systems,* 1828 (2015).

288. Sutskever, I., Vinyals, O. & Le, Q. Sequence to sequence learning with neural networks. *Advances in neural information processing systems,* 3104 (2014).

289. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* **11,** 3371–3408 (2010).

290. Kingma, D. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

291. Zhang, B., Xiong, D., su jinsong, j., Duan, H. & Zhang, M. *Variational neural machine translation* in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2016), 521–530.

292. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

293. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361** (1995).

294. Lu, H. *et al.* FDCNet: filtering deep convolutional network for marine organism classification. *Multimed Tools Appl* **77,** 1–14 (2017).

295. Srivastava, R., Greff, K. & Schmidhuber, J. Training very deep networks. *Advances in neural information processing systems,* 2377 (2015).

296. Huang, T. *et al.* MOST: most-similar ligand based approach to target prediction. *BMC Bioinformatics* **18,** 165 (2017).

297. Huang, G., Liu, Z., Maaten, L. v. d. & Weinberger, K. Q. *Densely connected convolutional networks* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), 2261–2269.

298. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13,** 281–305 (2012).

299. Snoek, J., Larochelle, H. & Adams, R. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems,* 2951 (2012).

300. Snoek, J. *et al.* Scalable bayesian optimization using deep neural networks. *International conference on machine learning,* 2171 (2015).

301. Suganuma, M., Shirakawa, S. & Nagao, T. *A genetic programming approach to designing convolutional neural network architectures* in *Proceedings of the Genetic and Evolutionary Computation Conference on - GECCO '17* (ACM Press, New York, New York, USA, 2017), 497–504.

302. Sabour, S., Frosst, N. & Hinton, G. Dynamic routing between capsules. *Advances in neural information processing systems,* 3856.

303. Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. Convolutional sequence to sequence learning. *Proceedings of the 34th International Conference on Machine Learning-Volume 70,* 1243 (2017).

304. Kalchbrenner, N. *et al.* Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099* (2016).

305. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

306. Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

307. Luong, T., Pham, H. & Manning, C. D. *Effective Approaches to Attention-based Neural Machine Translation* in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2015), 1412–1421.

308. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems,* 5998 (2017).

309. Cho, K. *et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014), 1724–1734.

310. Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. *On the properties of neural machine translation: encoder–decoder approaches* in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014), 103–111.

311. Wu, Y. *et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

312. Schneider, G. Mind and machine in drug design. *Nat Mach Intell* **1,** 128–130 (2019).

313. Yu, L., Wang, S. & Lai, K. An integrated data preparation scheme for neural network data analysis. *IEEE Trans Knowl Data Eng* **18,** 217–230 (2006).

314. Lo, Y.-C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* **23,** 1538–1546 (2018).

315. *Handbook of molecular descriptors* (eds Todeschini, R. & Consonni, V.) (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2000).

316. Unterthiner, T., Ceulemans, H. & Steijaert, M. *Multi-task deep networks for drug target prediction* in *Advances in Neural Information Processing Systems* (2014), 1–4.

317. Rupp, M., Proschak, E. & Schneider, G. Kernel approach to molecular similarity based on iterative graph similarity. *J Chem Inf Model* **47,** 2280–2286 (2007).

318. Duvenaud, D. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems,* 2224 (2015).

319. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **4,** 268–276 (2018).

320. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. Application of generative autoencoder in de novo molecular design. *Mol Inform* **37** (2018).

321. Segall, M. D., Yusof, I. & Champness, E. J. Avoiding missed opportunities by analyzing the sensitivity of our decisions. *J Med Chem* **59,** 4267–4277 (2016).

322. Schneider, G. & Fechner, U. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* **4,** 649–663 (2005).

323. Schneider, G. Future de novo drug design. *Mol Inform* **33,** 397–402 (2014).

324. Böhm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* **6,** 61–78 (1992).

325. Gillet, V. J. *et al.* SPROUT: recent developments in the de novo design of molecules. *J Chem Inf Comput Sci* **34,** 207–217 (1994).

326. Ruddigkeit, L., Blum, L. C. & Reymond, J.-L. Visualization and virtual screening of the chemical universe database GDB-17. *J Chem Inf Model* **53,** 56–65 (2013).

327. Hartenfeller, M. *et al.* DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol* **8,** e1002380 (2012).

328. Schneider, P. *et al.* Target Profile Prediction and Practical Evaluation of a Biginelli-Type Dihydropyrimidine Compound Library. *Pharmaceuticals* **4,** 1236–1247 (2011).

329. Hartenfeller, M. & Schneider, G. Enabling future drug discovery by de novo design. *WIREs Comput Mol Sci* **1,** 742–759 (2011).

330. Friedrich, L., Rodrigues, T., Neuhaus, C. S., Schneider, P. & Schneider, G. From complex natural products to simple synthetic mimetics by computational de novo design. *Angew Chem Int Ed Engl* **55,** 6789–6792 (2016).

331. Besnard, J. *et al.* Automated design of ligands to polypharmacological profiles. *Nature* **492,** 215–220 (2012).

332. Ikebata, H., Hongo, K., Isomura, T., Maezono, R. & Yoshida, R. Bayesian molecular design with a chemical language model. *J Comput Aided Mol Des* **31,** 379–391 (2017).

333. Miyao, T., Kaneko, H. & Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J Chem Inf Model* **56,** 286–299 (2016).

334. Churchwell, C. J. *et al.* The signature molecular descriptor. 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. *J Mol Graph Model* **22,** 263–273 (2004).

335. Wong, W. W. & Burkowski, F. J. A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem. *J Cheminform* **1,** 4 (2009).

336. Miyao, T., Arakawa, M. & Funatsu, K. Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Mol Inform* **29,** 111–125 (2010).

337. Takeda, S., Kaneko, H. & Funatsu, K. Chemical-Space-Based de Novo Design Method To Generate Drug-Like Molecules. *J Chem Inf Model* **56,** 1885–1893 (2016).

338. Mishima, K., Kaneko, H. & Funatsu, K. Development of a new de novo design algorithm for exploring chemical space. *Mol Inform* **33,** 779–789 (2014).

339. Kawai, K., Nagata, N. & Takahashi, Y. De novo design of drug-like molecules by a fragment-based molecular evolutionary approach. *J Chem Inf Model* **54,** 49–56 (2014).

340. Dey, F. & Caflisch, A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J Chem Inf Model* **48,** 679–690 (2008).

341. Fechner, U. & Schneider, G. Flux (1): a virtual synthesis scheme for fragment-based de novo design. *J Chem Inf Model* **46,** 699–707 (2006).

342. Schneider, G. Generative Models for Artificially-intelligent Molecular Design. *Mol Inform* **37** (2018).

343. Kadurin, A. *et al.* The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **8,** 10883–10890 (2017).

344. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol Pharm* **14,** 3098–3104 (2017).

345. Jaques, N. *et al.* Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. *Proceedings of the 34th International Conference on Machine Learning-Volume 70,* 1645 (2017).

346. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J Cheminform* **9,** 48 (2017).

347. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent Sci* **4,** 120–131 (2018).

348. Yu, L., Zhang, W., Wang, J. & Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

349. Harel, S. & Radinsky, K. Prototype-Based Compound Discovery Using Deep Generative Models. *Mol Pharm* **15,** 4406–4416 (2018).

350. Yuan, W. *et al.* Chemical space mimicry for drug discovery. *J Chem Inf Model* **57,** 875–882 (2017).

351. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun Chem* **1,** 68 (2018).

352. Arús-Pous, J. *et al.* Exploring the GDB-13 chemical space using deep generative models. *J Cheminform* **11,** 20 (2019).

353. Pogany, P., Arad, N., Genway, S. & Pickett, S. D. De novo Molecule Design by Translating from Reduced Graphs to SMILES. *J Chem Inf Model* (2018).

354. Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform* **10,** 31 (2018).

355. Awale, M., Sirockin, F., Stiefl, N. & Reymond, J.-L. Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks. *J Chem Inf Model* **59,** 1347–1356 (2019).

356. Gupta, A. *et al.* Generative recurrent networks for de novo drug design. *Mol Inform* **37** (2018).

357. Skalic, M., Jiménez, J., Sabbadin, D. & De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J Chem Inf Model* **59,** 1205–1214 (2019).

358. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J Chem Inf Model* **57,** 942–957 (2017).

359. Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat Biotechnol* **27,** 157–167 (2009).

360. Hopkins, A. L. Drug discovery: Predicting promiscuity. *Nature* **462,** 167–168 (2009).

361. Van Rossum, G. *The Python Language Reference —Python 2.7.13 documentation* 2017.

362. Decherchi, S. & Rocchia, W. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PLoS ONE* **8,** e59744 (2013).

363. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38,** 305–320 (1996).

364. Theil, H. *A rank-invariant method of linear and polynomial regression analysis (III)* in *Proceedings of the KNAW* **53** (Noord-Hollandsche Uitgevers Maatschappij, Amsterdam, 1950), 1397–1412.

365. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* **39,** D411–9 (2011).

366. Richards, F. M. Areas, volumes, packing and protein structure. *Annual review of biophysics and bioengineering* **6,** 151–176 (1977).

367. Adamczak, R., Porollo, A. & Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* **59,** 467–475 (2005).

368. Richmond, T. J. Solvent accessible surface area and excluded volume in proteins. *Journal of Molecular Biology* **178,** 63–89 (1984).

369. Jackson, R. M. & Sternberg, M. J. E. Protein surface area defined. *Nature* **366,** 638–638 (1993).

370. Jackson, R. M. & Sternberg, M. J. Application of scaled particle theory to model the hydrophobic effect: implications for molecular association and protein stability. *Protein Engineering* **7,** 371–383 (1994).

371. Krack, M. & Jug, K. in *Molecular Electrostatic Potentials - Concepts and Applications* 297–331 (Elsevier, 1996).

372. Lin, S. L., Nussinov, R., Fischer, D. & Wolfson, H. J. Molecular surface representations by sparse critical points. *Proteins* **18,** 94–101 (1994).

373. Cazals, F., Chazal, F. & Lewiner, T. *Molecular shape analysis based upon the morse-smale complex and the connolly function* in *Proceedings of the nineteenth conference on Computational geometry - SCG '03* (ACM Press, New York, New York, USA, 2003), 351.

374. Mitra, P. & Pal, D. New measures for estimating surface complementarity and packing at protein-protein interfaces. *FEBS Letters* **584,** 1163–1168 (2010).

375. Jiang, F. & Kim, S. H. 'Soft docking': matching of molecular surface cubes. *J Mol Biol* **219,** 79–102 (1991).

376. Allen, F. H., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. in *International Tables for Crystallography: Mathematical, physical and chemical tables* (ed Prince, E.) 790–811 (International Union of Crystallography, Chester, England, 2006).

377. Eastman, P. *et al.* Openmm 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J Chem Theory Comput* **9,** 461–469 (2013).

378. Kroemer, R. T. *et al.* Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J Chem Inf Comput Sci* **44,** 871–881 (2004).

379. Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* **320,** 597–608 (2002).

380. Hassell, A. M. *et al.* Crystallization of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* **63,** 72–79 (2007).

381. Le Maire, A. *et al.* In-plate protein crystallization, in situ ligand soaking and X-ray diffraction. *Acta Crystallogr D Biol Crystallogr* **67,** 747–755 (2011).

382. Sheridan, R. P., McGaughey, G. B. & Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J Comput Aided Mol Des* **22,** 257–265 (2008).

383. Ebejer, J.-P., Morris, G. M. & Deane, C. M. Freely available conformer generation methods: how good are they? *J Chem Inf Model* **52,** 1146–1158 (2012).

384. Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J Chem Inf Model* **55,** 2562–2574 (2015).

385. Boström, J., Greenwood, J. R. & Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* **21,** 449–462 (2003).

386. Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry* (1996).

387. Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J Comput Chem* **20,** 720–729 (1999).

388. Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J Comput Chem* **20,** 730–748 (1999).

389. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J Cheminform* **6,** 37 (2014).

390. Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A. & Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* **50,** 572–584 (2010).

391. Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor Chim Acta* **28,** 213–222 (1973).

392. Theiler, J. Efficient algorithm for estimating the correlation dimension from a set of discrete points. *Phys. Rev. A* **36,** 4456–4462 (1987).

393. Rousseeuw, P. J. & Leroy, A. M. *Robust regression and outlier detection* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1987).

394. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J Chem Doc* **5,** 107–113 (1965).

395. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Model* **29,** 97–101 (1989).

396. Tiikkainen, P. *et al.* Critical comparison of virtual screening methods against the MUV data set. *J Chem Inf Model* **49,** 2168–2178 (2009).

397. Vogel, S. M., Bauer, M. R. & Boeckler, F. M. DEKOIS: demanding evaluation kits for objective in silico screening–a versatile tool for benchmarking docking programs and scoring functions. *J Chem Inf Model* **51,** 2650–2665 (2011).

398. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **42,** D1083–90 (2014).

399. Gamo, F.-J. *et al.* Thousands of chemical starting points for antimalarial lead identification. *Nature* **465,** 305–310 (2010).

400. Todeschini, R. *et al.* Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J Chem Inf Model* **52,** 2884–2901 (2012).

401. Muchmore, S. W. *et al.* Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J Chem Inf Model* **48,** 941–948 (2008).

402. Marschner, I. C. & Gillett, A. C. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* **13,** 179–192 (2012).

403. Friedrich, L. *et al.* Shape Similarity by Fractal Dimensionality: An Application in the de novo Design of (-)-Englerin A Mimetics. *ChemMedChem* (2020).

404. Carson, C. *et al.* Englerin A agonizes the TRPC4/C5 cation channels to inhibit tumor cell line proliferation. *PLoS ONE* **10,** e0127498 (2015).

405. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci USA* **111,** 4067–4072 (2014).

406. Reker, D. *et al.* Revealing the macromolecular targets of complex natural products. *Nat Chem* **6,** 1072–1078 (2014).

407. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform* **1,** 8 (2009).

408. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* **29,** 1046–1051 (2011).

409. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol* **26,** 127–132 (2008).

410. Davies, M. *et al.* ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* **43,** W612–20 (2015).

411. Bosc, N., Meyer, C. & Bonnet, P. The use of novel selectivity metrics in kinase research. *BMC Bioinformatics* **18,** 17 (2017).

412. Graczyk, P. P. Gini coefficient: a new way to express selectivity of kinase inhibitors against a family of kinases. *J Med Chem* **50,** 5773–5779 (2007).

413. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48,** 443–453 (1970).

414. Kalliokoski, T., Ronkko, T. P. & Poso, A. Increasing the Throughput of Shape-Based Virtual Screening with GPU Processing and Single Conformation Databases. *Mol Inform* **29,** 293–296 (2010).

415. Vogt, M., Stumpfe, D., Geppert, H. & Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J Med Chem* **53,** 5707–5715 (2010).

416. Nicholls, A. *et al.* Molecular shape and medicinal chemistry: a perspective. *J Med Chem* **53,** 3862–3886 (2010).

417. Nicholls, A., MacCuish, N. E. & MacCuish, J. D. Variable selection and model validation of 2D and 3D molecular descriptors. *J Comput Aided Mol Des* **18,** 451–474 (2004).

418. Hawkins, P. C. D. & Nicholls, A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J Chem Inf Model* **52,** 2919–2936 (2012).

419. Kirchmair, J., Laggner, C., Wolber, G. & Langer, T. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J Chem Inf Model* **45,** 422–430 (2005).

420. Kirchmair, J., Wolber, G., Laggner, C. & Langer, T. Comparative performance assessment of the conformational model generators omega and catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations. *J Chem Inf Model* **46,** 1848–1861 (2006).

421. D'Agostino, R. B. An omnibus test of normality for moderate and large size samples. *Biometrika* **58,** 341–348 (1971).

422. D'Agostino, R. B. & Pearson, E. S. Tests for Departure from Normality. Empirical Results for the Distributions of b^2 and $\sqrt{b^1}$. *Biometrika* **60,** 613 (1973).

423. Lopes, J. C. D., Dos Santos, F. M., Martins-José, A., Augustyns, K. & De Winter, H. The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *J Cheminform* **9,** 7 (2017).

424. Ballester, P. J. & Richards, W. G. Ultrafast shape recognition for similarity search in molecular databases. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **463,** 1307–1321 (2007).

425. Haque, I. S., Pande, V. S. & Walters, W. P. Anatomy of high-performance 2D similarity calculations. *J Chem Inf Model* **51,** 2345–2351 (2011).

426. Haque, I. S. & Pande, V. S. SCISSORS: a linear-algebraical technique to rapidly approximate chemical similarities. *J Chem Inf Model* **50,** 1075–1088 (2010).

427. Haque, I. S. & Pande, V. S. Error bounds on the SCISSORS approximation method. *J Chem Inf Model* **51,** 2248–2253 (2011).

428. Eckert, H. & Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* **12,** 225–233 (2007).

429. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71,** 58–63 (2015).

430. Maggiora, G. M. On outliers and activity cliffs–why QSAR often disappoints. *J Chem Inf Model* **46,** 1535 (2006).

431. Guha, R. & Van Drie, J. H. Structure–activity landscape index: identifying and quantifying activity cliffs. *J Chem Inf Model* **48,** 646–658 (2008).

432. Jaakola, V.-P. *et al.* The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **322,** 1211–1217 (2008).

433. Diniz, C. *et al.* Ligands and therapeutic perspectives of adenosine A(2A) receptors. *Curr Pharm Des* **14,** 1698–1722 (2008).

434. Leone, R. D., Lo, Y.-C. & Powell, J. D. A2aR antagonists: Next generation checkpoint blockade for cancer immunotherapy. *Comput Struct Biotechnol J* **13,** 265–272 (2015).

435. Christensen, R., Kristensen, P. K., Bartels, E. M., Bliddal, H. & Astrup, A. Efficacy and safety of the weight-loss drug rimonabant: a meta-analysis of randomised trials. *Lancet* **370,** 1706–1713 (2007).

436. Meiri, E. *et al.* Efficacy of dronabinol alone and in combination with ondansetron versus ondansetron alone for delayed chemotherapy-induced nausea and vomiting. *Curr Med Res Opin* **23,** 533–543 (2007).

437. Li, X. *et al.* mGluR5 antagonism inhibits cocaine reinforcement and relapse by elevation of extracellular glutamate in the nucleus accumbens via a CB1 receptor mechanism. *Sci Rep* **8,** 3686 (2018).

438. Davis, M. T. *et al.* In vivo evidence for dysregulation of mGluR5 as a biomarker of suicidal ideation. *Proc Natl Acad Sci USA* **116,** 11490–11495 (2019).

439. Stahn, C., Löwenberg, M., Hommes, D. W. & Buttgereit, F. Molecular mechanisms of glucocorticoid action and selective glucocorticoid receptor agonists. *Mol Cell Endocrinol* **275,** 71–78 (2007).

440. Spitz, I. M. & Bardin, C. W. Mifepristone (RU 486)–a modulator of progestin and glucocorticoid action. *N Engl J Med* **329,** 404–412 (1993).

441. Fleseriu, M. *et al.* Mifepristone, a glucocorticoid receptor antagonist, produces clinical and metabolic benefits in patients with Cushing's syndrome. *J Clin Endocrinol Metab* **97,** 2039–2049 (2012).

442. Tachibana, K., Yamasaki, D., Ishimoto, K. & Doi, T. The role of ppars in cancer. *PPAR Research* **2008,** 102737 (2008).

443. Lamers, C., Schubert-Zsilavecz, M. & Merk, D. Therapeutic modulators of peroxisome proliferator-activated receptors (PPAR): a patent review (2008-present). *Expert opinion on therapeutic patents* **22,** 803–841 (2012).

444. Merk, D. *et al.* Anthranilic acid derivatives as nuclear receptor modulators–development of novel PPAR selective and dual PPAR/FXR ligands. *Bioorganic & Medicinal Chemistry* **23,** 499–514 (2015).

445. Gupta, R. A. *et al.* Prostacyclin-mediated activation of peroxisome proliferator-activated receptor delta in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 13275–13280 (2000).

446. Takayama, O. *et al.* Expression of PPARdelta in multistage carcinogenesis of the colorectum: implications of malignant cancer morphology. *British Journal of Cancer* **95,** 889–895 (2006).

447. Liu, Y. *et al.* The Role of PPAR-$\delta inMetabolism, Inflammation, andCancer : ManyCharactersofaCriticalTranscriptionFactor.$. *International Journal of Molecular Sciences* **19** (2018).

448. Chen, Y. R. & Tan, T. H. Inhibition of the c-Jun N-terminal kinase (JNK) signaling pathway by curcumin. *Oncogene* **17,** 173–178 (1998).

449. Bennett, B. L. *et al.* SP600125, an anthrapyrazolone inhibitor of Jun N-terminal kinase. *Proc Natl Acad Sci USA* **98,** 13681–13686 (2001).

450. Borsello, T. *et al.* A peptide inhibitor of c-Jun N-terminal kinase protects against excitotoxicity and cerebral ischemia. *Nat Med* **9,** 1180–1186 (2003).

451. Velazquez, R., Shaw, D. M., Caccamo, A. & Oddo, S. Pim1 inhibition as a novel therapeutic strategy for Alzheimer's disease. *Mol Neurodegener* **11,** 52 (2016).

452. Bachmann, M. & Möröy, T. The serine/threonine kinase Pim-1. *Int J Biochem Cell Biol* **37,** 726–730 (2005).

453. Foulks, J. M. *et al.* A small-molecule inhibitor of PIM kinases as a potential treatment for urothelial carcinomas. *Neoplasia* **16,** 403–412 (2014).

454. Arunesh, G. M., Shanthi, E., Krishna, M. H., Sooriya Kumar, J. & Viswanadhan, V. N. Small molecule inhibitors of PIM1 kinase: July 2009 to February 2013 patent update. *Expert Opin Ther Pat* **24,** 5–17 (2014).

455. Keeton, E. K. *et al.* AZD1208, a potent and selective pan-Pim kinase inhibitor, demonstrates efficacy in preclinical models of acute myeloid leukemia. *Blood* **123,** 905–913 (2014).

456. Morwick, T. Pim kinase inhibitors: a survey of the patent literature. *Expert Opin Ther Pat* **20,** 193–212 (2010).

457. Cheeseright, T. J., Mackey, M. D., Melville, J. L. & Vinter, J. G. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *J Chem Inf Model* **48,** 2108–2117 (2008).

458. Hu, Y., Stumpfe, D. & Bajorath, J. Recent advances in scaffold hopping. *Journal of Medicinal Chemistry* **60,** 1238–1246 (2017).

459. Johnson, M. & Maggiora, G. Concepts and applications of molecular similarity. *Concepts and applications of molecular similarity* (1990).

460. Fontaine, F., Bolton, E., Borodina, Y. & Bryant, S. H. Fast 3D shape screening of large chemical databases through alignment-recycling. *Chem Cent J* **1,** 12 (2007).

461. Bullock, A. N. *et al.* Crystal structure of the PIM2 kinase in complex with an organoruthenium inhibitor. *PLoS ONE* **4,** e7112 (2009).

462. Wang, H.-L., Cee, V. J., Herberich, B. J. & Jackson, C. L. Patent US:9321756 (2016).

463. Mikkers, H. *et al.* Mice deficient for all PIM kinases display reduced body size and impaired responses to hematopoietic growth factors. *Mol Cell Biol* **24,** 6104–6115 (2004).

464. Kantarjian, H. M. *et al.* Nilotinib (formerly AMN107), a highly selective BCR-ABL tyrosine kinase inhibitor, is effective in patients with Philadelphia chromosome-positive chronic myelogenous leukemia in chronic phase following imatinib resistance and intolerance. *Blood* **110,** 3540–3546 (2007).

465. Manley, P. W. *et al.* Imatinib: a selective tyrosine kinase inhibitor. *European Journal of Cancer* **38 Suppl 5,** S19–27 (2002).

466. Eid, S., Turk, S., Volkamer, A., Rippmann, F. & Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics* **18,** 16 (2017).

467. Debreczeni, J. E. *et al.* Ruthenium half-sandwich complexes bound to protein kinase Pim-1. *Angew Chem Int Ed Engl* **45,** 1580–1585 (2006).

468. Jope, R. S., Yuskaitis, C. J. & Beurel, E. Glycogen synthase kinase-3 (GSK3): inflammation, diseases, and therapeutics. *Neurochem Res* **32,** 577–595 (2007).

469. Wang, Z. *et al.* Glycogen synthase kinase 3 in MLL leukaemia maintenance and targeted therapy. *Nature* **455,** 1205–1209 (2008).

470. Yoshino, Y. & Ishioka, C. Inhibition of glycogen synthase kinase-3 beta induces apoptosis and mitotic catastrophe by disrupting centrosome regulation in cancer cells. *Sci Rep* **5,** 13249 (2015).

471. Park, Y. *et al.* Cytoplasmic DRAK1 overexpressed in head and neck cancers inhibits TGF-β tumor suppressor activity by binding to Smad3 to interrupt its complex formation with Smad4. *Oncogene* **34,** 5037–5045 (2015).

472. Ranek, M. J. *et al.* PKG1-modified TSC2 regulates mTORC1 activity to counter adverse cardiac stress. *Nature* **566,** 264–269 (2019).

473. Ingley, E. Functions of the Lyn tyrosine kinase in health and disease. *Cell Commun Signal* **10,** 21 (2012).

474. Pierce, A. C., Jacobs, M. & Stuver-Moody, C. Docking study yields four novel inhibitors of the protooncogene Pim-1 kinase. *J Med Chem* **51,** 1972–1975 (2008).

475. Jacobs, M. D. *et al.* Pim-1 ligand-bound structures reveal the mechanism of serine/threonine kinase inhibition by LY294002. *The Journal of Biological Chemistry* **280,** 13728–13734 (2005).

476. Qian, K. C. *et al.* Structural basis of constitutive activity and a unique nucleotide binding mode of human Pim-1 kinase. *The Journal of Biological Chemistry* **280,** 6130–6137 (2005).

477. Bosshard, H. R., Marti, D. N. & Jelesarov, I. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *Journal of Molecular Recognition* **17,** 1–16 (2004).

478. Carugo, O. How large B-factors can be in protein crystal structures. *BMC Bioinformatics* **19,** 61 (2018).

479. LLC, S. *Schrödinger Release 2019-2: Maestro* New York, 2019.

480. Tsuganezawa, K. *et al.* A novel Pim-1 kinase inhibitor targeting residues that bind the substrate peptide. *Journal of Molecular Biology* **417,** 240–252 (2012).

481. Ren, J.-X. *et al.* Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking. *Journal of Chemical Information and Modeling* **51,** 1364–1375 (2011).

482. Kawasaki, Y. & Freire, E. Finding a better path to drug selectivity. *Drug Discovery Today* **16,** 985–990 (2011).

483. Hartenfeller, M., Proschak, E., Schüller, A. & Schneider, G. Concept of combinatorial de novo design of drug-like molecules by particle swarm optimization. *Chemical Biology & Drug Design* **72,** 16–26 (2008).

484. Boström, J., Hogner, A. & Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J Med Chem* **49,** 6716–6725 (2006).

485. Boström, J., Norrby, P. O. & Liljefors, T. Conformational energy penalties of protein-bound ligands. *J Comput Aided Mol Des* **12,** 383–396 (1998).

486. Boström, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J Comput Aided Mol Des* **15,** 1137–1152 (2001).

487. Melis, G., Dyer, C. & Blunsom, P. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589* (2017).

488. Popel, M. & Bojar, O. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics* **110** (2018).

489. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *Journal of Molecular Biology* **234,** 946–950 (1993).

490. Shin, W.-H., Zhu, X., Bures, M. G. & Kihara, D. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules* **20,** 12841–12862 (2015).

491. Voet, A., Berenger, F. & Zhang, K. Y. J. Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors. *PLoS ONE* **8,** e75762 (2013).

492. Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem Biol* **3,** 973–980 (1996).

493. Li, Z. & Lazaridis, T. Thermodynamics of buried water clusters at a protein-ligand binding interface. *J Phys Chem B* **110,** 1464–1475 (2006).

494. Li, Z. & Lazaridis, T. Water at biomolecular binding interfaces. *Phys Chem Chem Phys* **9,** 573–581 (2007).

495. Rodier, F., Bahadur, R. P., Chakrabarti, P. & Janin, J. Hydration of protein-protein interfaces. *Proteins* **60,** 36–45 (2005).

496. Lu, Y., Wang, R., Yang, C.-Y. & Wang, S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J Chem Inf Model* **47,** 668–675 (2007).

497. Schwaller, P. *et al.* Molecular Transformer – A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ChemrXiv preprint ChemrXiv:10.26434/chemrxiv.7297379.v2* (2019).

498. Skalic, M., Varela-Rial, A., Jiménez, J., Martínez-Rosell, G. & De Fabritiis, G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. *Bioinformatics* **35,** 243–250 (2019).

499. Lovering, F. Escape from Flatland 2: complexity and promiscuity. *MedChemComm* **4,** 515–519 (2013).

500. Keiser, M. J., Irwin, J. J. & Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry* **49,** 10267–10276 (2010).

501. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25,** 25–29 (2000).

502. Consortium, G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43,** D1049–56 (2015).

503. Jenkins, J. L., Glick, M. & Davies, J. W. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J Med Chem* **47,** 6144–6159 (2004).

504. Zhang, Q. C., Petrey, D., Norel, R. & Honig, B. H. Protein interface conservation across structure space. *Proc Natl Acad Sci USA* **107,** 10896–10901 (2010).

505. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* **5,** e1000585 (2009).

Appendix

Englerin study

Parts of this section are published as: Shape Similarity by Fractal Dimensionality: An Application in the de novo Design of (-)-Englerin A Mimetics. [403]
Authors: Lukas Friedrich, Ryan Byrne, Michael Mederos y Schnitzler, Aaron Treder, Inderjeet Singh, Christoph Bauer, Thomas Gudermann, Ursula Storch and Gisbert Schneider

Predicted active compounds

Physicochemical properties

Bioactivity determination

In vitro biological assessment Intracellular Calcium Assays The modulatory effects of compounds 1 and 2 were tested in a cell-based intracellular calcium assay for transient receptor potential melastatin 8 cation channel (TRPM8). Log(concentration) response curves (four-parameter logistic curves) were plotted in Prism 7 (GraphPad Software, La Jolla, CA, USA). IC_{50} values were determined with Prism 7. Dissociation constants (Ki) were calculated with the modified Cheng-Prusoff equation:

$$K_i = IC_{50}\left[1 + \frac{C}{EC_{50} \cdot C}\right]^{-1} \qquad (.1)$$

where C is the concentration of control activator icilin (0.1 µM) in the assay and EC_{50}.C its EC_{50} value (0.016 µM). Modulatory effects of com-

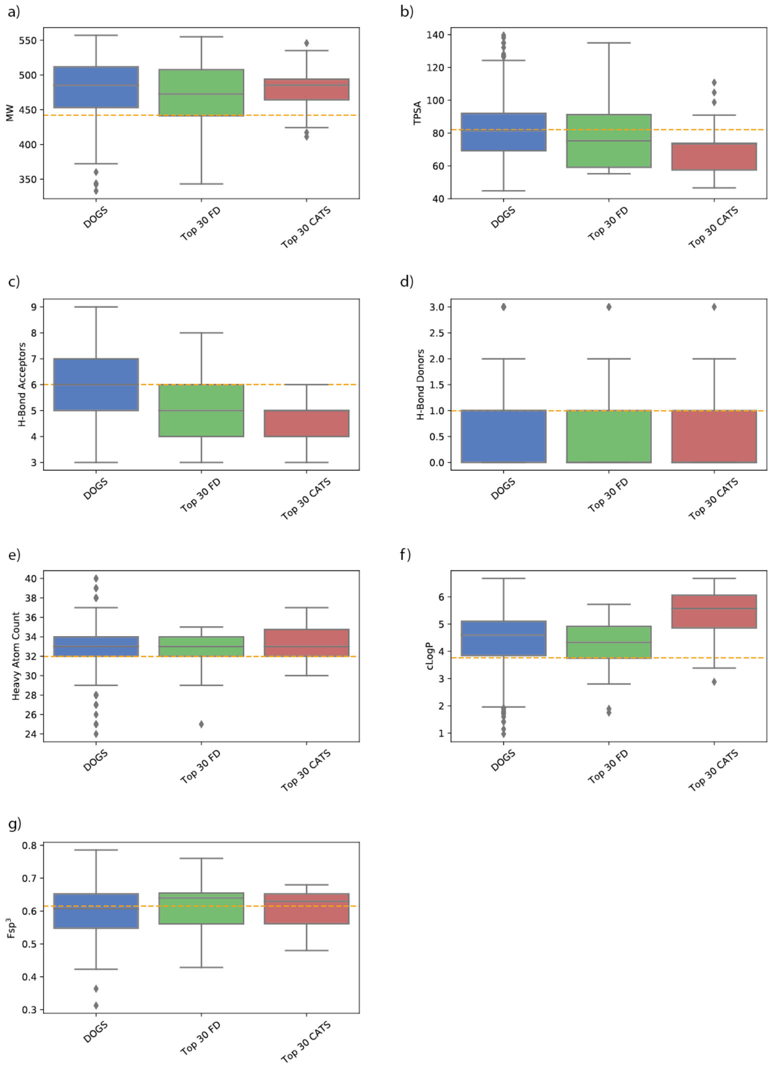


Figure 1: Calculated physicochemical properties of the DOGS set (blue), TOP30 FD (green), TOP30 CATS (red) and (-)-Englerin A (dashed, orange line). Box-Whisker plots of the a) molecular weight, b) topological polar surface area (TPSA), c) number of hydrogen-bond acceptors, d) number of hydrogen-bond donors, e) number of non-hydrogen atoms, f) calculated octanol-water partition coefficient (cLogP), g) fraction of sp3-hybridized carbon atoms.

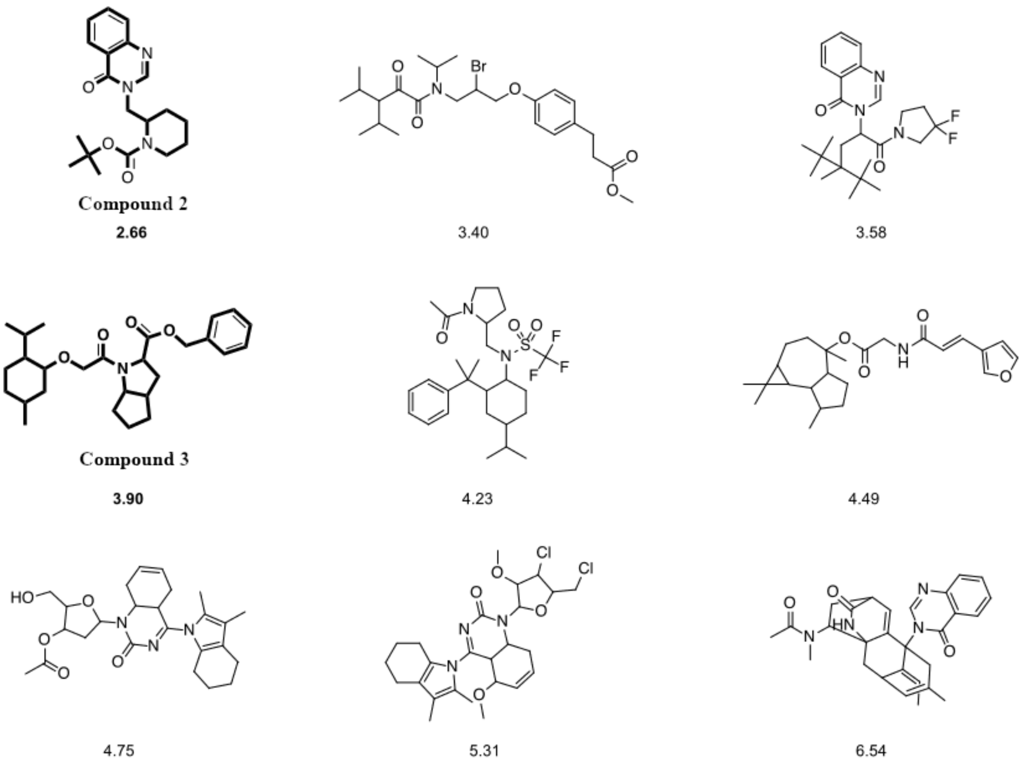


Figure 2: The top-30 ranked compounds according to GFD towards **1a**. Compounds selected for synthesis and testing (Compounds **2** and **3**) are highlighted. For each entry, the synthesisability score is included. Compounds were chosen by a medicinal chemist based on low synthesisability score and ready availability of building blocks

pounds 1 and 2 on TRPM8 were determined by intracellular calcium response detected by fluorimetry. Inhibitory effects of 1 and 2 on TRPA1, TRPV3 and TRPV4 were measured in calcium assays, performed on a Molecular Devices' FLIPRTETRA (fluorescent imaging plate reader). Allyl isothiocyanate (AITC, EC80 = 10 µM, TRPA1), 2-aminoethoxydiphenyl borate (2-APB, EC80 = 50 µM, TRPV3), and GSK106790A (EC80 = 100 nM, TRPV4) as reference activators. Ruthenium Red was used as reference inhibitor in TRPA1, TRPV3, and TRPV4 assays. Calcium assays (TRPA1, TRPV3, TRPV4, and TRPM8) were conducted by Eurofins Cerep SA (France) and Eurofins Panlabs (USA) on a fee-for-service basis.

Electrophysiology Human embryonic kidney (HEK293) cells (293T, ATCC CRL-3216) were maintained in Earl's MEM (Sigma-Aldrich, Taufkirchen, Germany), with 100 units ml^{-1} penicillin and 100 µg ml^{-1} streptomycin supplemented with 10% (vol/vol) FCS (Gibco, Thermo Fisher Scientific, Waltham, MA, USA) and 2 mM glutamine. All cells were held at 37○C in a humidified atmosphere with 5% CO2. Cells were seeded into 6-well dishes and transiently transfected at confluency of about 90% using GeneJuice (Merck Millipore, Billerica, MA, USA) according to the manufacturer's protocol. Conventional whole-cell recordings were carried out at room temperature 15 hours after transfection with the human TRPM8 (NP_076985) in pCAGGSM2-IRES-GFP expression vector or 24 hours after transfection of the cells with the rat TRPC4 isoform TRPC4-beta1 (NP_001076584) in pIRES2-eGFP expression vector. The following bath solution containing 140 mM NaCl, 5 mM CsCl, 1 mM MgCl2, 2 mM CaCl2, 10 mM glucose, 10 mM HEPES (pH 7.4 with NaOH) and resulting in an osmolarity of 295-302 mOsm kg^{-1} was used. The pipette solution for TRPM8 measurements contained 130 mM CsCl, 5.792 mM MgCl2, 0.524 mM CaCl2, 10 mM BAPTA (5.5 nM free Ca2+), 1 mM HEDTA (3 mM free Mg2+) and 10 mM HEPES (pH 7.2 with CsOH), resulting in an osmolality of 296 mOsm kg^{-1}. The pipette solution for TRPC4 measurements contained 120 mM CsCl, 9.4 mM NaCl, 0.2 mM Na3-GTP, 1 mM MgCl2, 3.949 mM CaCl2, 10 mM BAPTA (100 nM free Ca2+) and 10 mM HEPES (pH 7.2 with CsOH), resulting in an osmolality of 296 mOsm kg^{-1}.

Patch pipettes made of borosilicate glass (Science Products, Hofheim, Germany) had resistances of 2.0-2.8 MΩ for the whole-cell measurements. Data were collected with an EPC10 patch clamp amplifier (HEKA, Lambrecht, Germany) using the Patchmaster software (HEKA). Current density-voltage relations were obtained from voltage ramps from –100 to +100 mV with a slope of 0.5 V s^{-1} applied at a frequency of 2 Hz. Data were acquired at a frequency of 5 kHz after filtering at 1.67 kHz. For TRPM8 channel activation, 200 µM (-)-menthol was applied. 0.1, 1 and 10 µM compound 2 was applied in the presence of (-)-menthol. In some measurements (-)-menthol and compound 2 were washed out and a second application of (-)-menthol caused second TRPM8 current increases. The maximal (-)-menthol-induced outward currents at +100 mV before application of compound 2 were used for analysis. TRPM8-expressing cells which showed basal activity $\geq$2 nA/pF at +100 mV were excluded from further analysis. To determine IC$_{50}$ value, 0.3, 1, 2, 3, 10, 30 and 100 µM compound 2 was applied. For TRPC4 channel activation 50 nM (-)-Englerin A (Carl Roth, Karlsruhe, Germany) was applied two times.

(-)-Englerin A was applied in the presence of different compound 1 and compound 2 concentrations. The second (-)-Englerin A-induced current increase was used for normalization. For calculation of IC_{50}, maximal (-)-Englerin A-induced outward currents at +100 mV were used. For calculation of the percentage of maximal outward currents at +100 mV basal currents before application of the first stimulus were always subtracted. Dissociation constants (Ki) were calculated with the modified Cheng Prusoff equation(.1), where C is the concentration of control activator (-)-Englerin A (0.05 µM) in the assay and EC_{50} its EC_{50} value (0.0112 µM).

Statistical analysis Electrophysiological data was analyzed using Origin 7.5 software (OriginLab, Northampton, MA, USA). Data are presented as mean ± standard error of the mean (s.e.m.). For calculation of IC_{50} value, concentration response curve was fitted using Single Hill-equation until no reduction of Chi-square was noted.

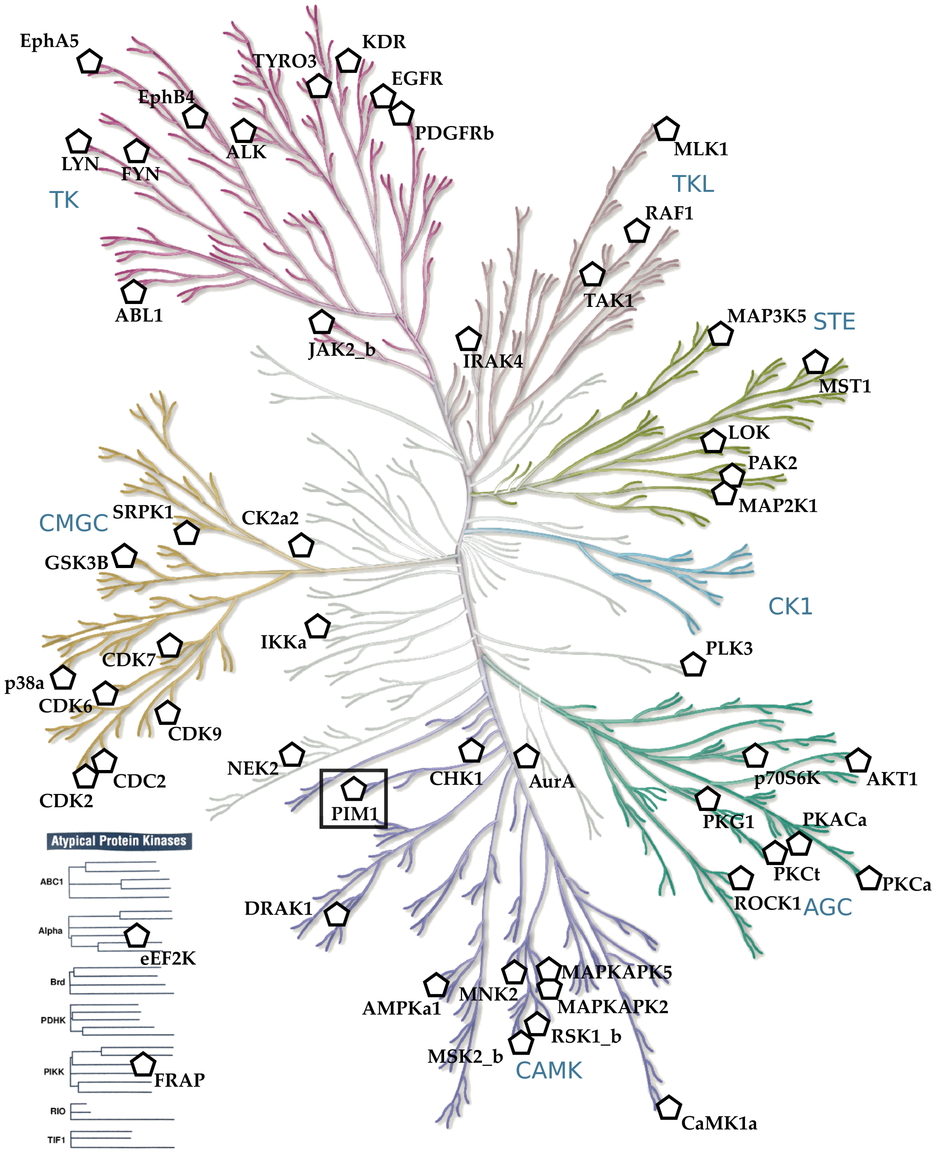
PIM1 Study

Kinase Selectivity

Table 1: Data obtained from fee-for-service provider Cerep Eurofins regarding kinase selectivity of compound **102** with regard a panel of medically-relevant kinases of interest in drug discovery.

Kinase	Compound 102% inhibition at 10 μM
Abl(h)	32
ALK(h)	7
AMPKα1(h)	27
ASK1(h)	0
Aurora-A(h)	21
CaMKI(h)	18
CDK1/cyclinB(h)	12
CDK2/cyclinA(h)	9
CDK6/cyclinD3(h)	20
CDK7/cyclinH/MAT1(h)	8
CDK9/cyclin T1(h)	22
CHK1(h)	56
CK1γ1(h)	62
CK2α2(h)	18
c-RAF(h)	37
DRAK1(h)	87
eEF-2K(h)	8
EGFR(h)	42
EphA5(h)	20
EphB4(h)	0
Fyn(h)	21
GSK3β(h)	82
IGF-1R(h)	0
IKKα(h)	0
IRAK4(h)	8
JAK2(h)	7
KDR(h)	0
LOK(h)	0
Lyn(h)	33
MAPKAP-K2(h)	16

MEK1(h)	21
MLK1(h)	40
Mnk2(h)	41
MSK2(h)	29
MST1(h)	15
mTOR(h)	15
NEK2(h)	16
p70S6K(h)	34
PAK2(h)	0
PDGFRβ(h)	13
PIM1(h)	85
PKA(h)	19
PKBα(h)	30
PKCα(h)	19
PKCθ(h)	0
PKG1α(h)	30
Plk3(h)	9
PRAK(h)	33
ROCK-I(h)	3
Rse(h)	16
Rsk1(h)	56
SAPK2a(h)	12
SRPK1(h)	33
TAK1(h)	18
PI3 Kinase (p110b/p85a)(h)	3
PI3 Kinase (p120g)(h)	13
PI3 Kinase (p110d/p85a)(h)	13
PI3 Kinase (p110a/p85a)(h)	6



"Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com)"

Figure 3: Kinase map showing labelling for targets. N.B. kinase inhibition data are available for 58/536 kinases visible. PIM1 kinase is highlighted. Illustration reproduced courtesy of Cell Signalling Technology, Inc. (www.cellsignal.com)

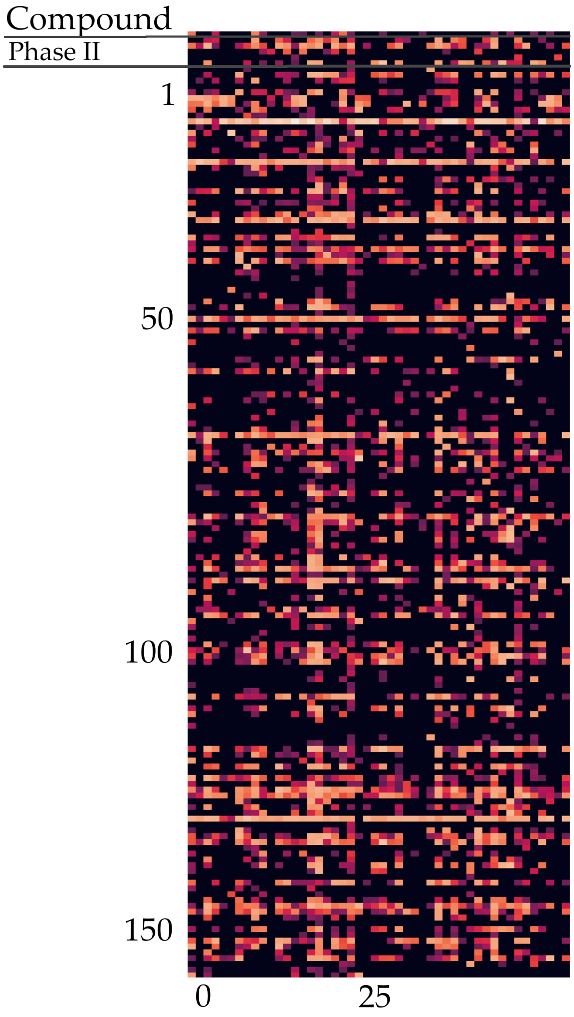


Figure 4: Heatmap of the kinase inhibition for each molecule studies in the kinase selectivity study, where brighter colours indicate higher inhibition at 10 µM. Compounds are ranked top-to-bottom by PIM1 inhibition, except for the first entry, which is for compound **102**, and the following five entries, which are for kinase inhibitors which have proceeded to at least phase II. Left-to-right order is by sequence similarity to the PIM1 Kinase.

Crystallographic parameters

Table 2: Crystallographic parameters obtained for the PIM1/Compound 911521 complex. Crystallisation was provided on a fee-for-service basis by SARomics Biostructures AB (Lund, Sweden). Figures in parentheses are from the highest resolution shell.

Parameter or Observation	**Value**
Protein Complex	PIM1/911521
Resolution (Å)	48.83 - 1.8 (1.84 - 1.80)
Wavelength (Å)	1.0332
Space group	P65

Unit cell (Å)	a = 97.65
	b = 97.65
	c = 80.51
	α = 90
	β = 90
	γ = 120

Completeness (%)	99.8 (98.0)
Redundancy	10.1 (9.7)
No. of observations / unique reflections	407886/22861
$<I/\sigma(I)>$	13.8 (1.51)
Rmerge (I) (%)	9.5 (174.2)
CC(1/2) (%)	99.8 (58.2)
Rmodel (F) (%)	15.7 (31.9)
Rfree (F) (%)	18.8 (38.7)
No. of non-hydrogen atoms	2601
No. of water molecules	283
RMS deviations from ideal geometry: Bond lengths (Å)	0.008
Bond angles (°)	1.5
Mean B-factor protein chain A ($Å^2$)	37.4
Mean B-factor ligands, chain B ($Å^2$)	34.3
Mean B-factor, Imidazole, chain C ($Å^2$)	42.8
Mean B-factor, PEG, chain D ($Å^2$)	53.3
Mean B-factor solvent, chain H ($Å^2$)	47.4
Ramachandran plot quality*	
Favoured regions (%)	97.8
Allowed regions (%)	2.2
Outliers (%)	0.0

Ryan Byrne

Personal Data

Name:	Ryan Byrne
email:	rbyrne944@gmail.com

Scientific Education

09/2019	Doctoral degree, Dr. sc. ETH Zürich ETH Zürich
05/2016–06/2019	PhD student at ETH Zürich. Supervisor: Professor Dr. Gisbert Schneider
09/2017–03/2018	Visiting researcher, AstraZeneca Mölndal. Supervisors: Drs. Hongming Chen, Ola Engkvist
9/2016	Master of Science (MSc.) Bioinformatics and Theoretical Systems Biology Imperial College London
09/2015 - 09/2016	Supervisors: Professors M. Sternberg, M. Stumpf
6/2015	Master of Pharmacy (MPharm) Queen's University, Belfast
09/2010-06/2015	Supervisor: Professor. I. Tikhonova

Work Experience

2019-2020	Pharmaceutical AI Lead - Rejuveron Life Sciences AG
2016	AI Engineer Want2BeThere Ltd.
2015-2016	Bioinformatician, Genomics of Drug Sensitivity in Cancer, Garnett Group, Wellcome Trust Sanger Institute

Publications

L. Friedrich, R.Byrne, et al. Shape Similarity by Fractal Dimensionality: An Application in the de novo Design of (-)-Englerin A Mimetics.
ChemMedChem, 15(7), 566-570, 2020

R. Fino, R.Byrne, et al. Introducing the CSP Analyzer: a Novel Machine Learning-based Application for Automated Analysis of two-dimensional NMR spectra in NMR Fragment-based Screening.
Computational and Structural Biotechnology Journal, 18, 603-611, 2020

X. Yang, Y. Wang. R. Byrne. G. Schneider, S. Yang. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery.
Chemical Reviews, 119(18), 10520-10594, 2019

R. Byrne, G. Schneider. In Silico Target Prediction for Small Molecules.
Systems Chemical Biology, 273-309, 2019

F. Grisoni, D. Merk, R. Byrne, G. Schneider. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation.
Scientific reports, 8(1), 16469, 2018

Scholarships and Certificates

2016-2019	Marie Curie Early Stage Researcher - Fellowship
2016	Diploma of Imperial College (DIC)