Methods and Principles in Medicinal Chemistry

Edited by R. Mannhold, H. Kubinyi, G. Folkers

Editorial Board
H. Buschmann, H. Timmerman, H. van de Waterbeemd, T. Wieland

Previous Volumes of this Series:

Hoffmann, Rémy D. / Gohier, Arnaud / Pospisil, Pavel (Eds.)
Data Mining in Drug Discovery
2014
ISBN: 978-3-527-32984-7
Vol. 57

Dömling, Alexander (Ed.)
Protein-Protein Interactions in Drug Discovery
2013
ISBN: 978-3-527-33107-9
Vol. 56

Kalgutkar, Amit S. / Dalvie, Deepak / Obach, R. Scott / Smith, Dennis A.
Reactive Drug Metabolites
2012
ISBN: 978-3-527-33085-0
Vol. 55

Brown, Nathan (Ed.)
Bioisosteres in Medicinal Chemistry
2012
ISBN: 978-3-527-33015-7
Vol. 54

Gohlke, Holger (Ed.)
Protein-Ligand Interactions
2012
ISBN: 978-3-527-32966-3
Vol. 53

Kappe, C. Oliver / Stadler, Alexander / Dallinger, Doris
Microwaves in Organic and Medicinal Chemistry
Second, Completely Revised and Enlarged Edition
2012
ISBN: 978-3-527-33185-7
Vol. 52

Smith, Dennis A. / Allerton, Charlotte / Kalgutkar, Amit S. / van de Waterbeemd, Han / Walker, Don K.
Pharmacokinetics and Metabolism in Drug Design
Third, Revised and Updated Edition
2012
ISBN: 978-3-527-32954-0
Vol. 51

De Clercq, Erik (Ed.)
Antiviral Drug Strategies
2011
ISBN: 978-3-527-32696-9
Vol. 50

Klebl, Bert / Müller, Gerhard / Hamacher, Michael (Eds.)
Protein Kinases as Drug Targets
2011
ISBN: 978-3-527-31790-5
Vol. 49

Sotriffer, Christoph (Ed.)
Virtual Screening Principles, Challenges, and Practical Guidelines
2011
ISBN: 978-3-527-32636-5
Vol. 48
Scaffold Hopping in Medicinal Chemistry
Contents

List of Contributors XIII
Preface XVII
A Personal Foreword XIX

Part One Scaffolds: Identification, Representation Diversity, and Navigation 1

1 Identifying and Representing Scaffolds 3
Nathan Brown
1.1 Introduction 3
1.2 History of Scaffold Representations 4
1.3 Functional versus Structural Molecular Scaffolds 7
1.4 Objective and Invariant Scaffold Representations 7
1.4.1 Molecular Frameworks 7
1.4.2 Scaffold Tree 8
1.5 Maximum Common Substructures 9
1.6 Privileged Scaffolds 11
1.7 Conclusions 11
References 12

2 Markush Structures and Chemical Patents 15
David Anthony Cosgrove
2.1 Introduction 15
2.2 Encoding Markush Structures 18
2.2.1 The r_group Record 19
2.2.1.1 Exact R Groups 19
2.2.1.2 Inexact R Groups 19
2.2.1.3 Fused R Groups 19
2.2.2 The Menguin Program 20
2.2.3 Correspondence between the MIL File and the Markush Structure 21
2.3 The Search Algorithm 22
2.3.1 Matching R Groups 25
3 Scaffold Diversity in Medicinal Chemistry Space 39
Sarah R. Langdon, Julian Blagg, and Nathan Brown
3.1 Introduction 39
3.1.1 Scaffold Representation 39
3.1.2 What Do We Mean by Scaffold Diversity? 41
3.2 Scaffold Composition of Medicinal Chemistry Space 41
3.2.1 Natural Products as a Source of Novel Medicinal Chemistry Scaffolds 45
3.2.2 Enumerating Potential Medicinal Chemistry Scaffolds 46
3.2.3 Using Scaffold Composition to Interpret Bioactivity Data 48
3.3 Metrics for Quantifying the Scaffold Diversity of Medicinal Chemistry Space 48
3.4 Visualizing the Scaffold Diversity of Medicinal Chemistry Space 53
3.5 Conclusions 56
References 57

4 Scaffold Mining of Publicly Available Compound Data 61
Ye Hu and Jürgen Bajorath
4.1 Introduction 61
4.2 Scaffold Definition 62
4.3 Selectivity of Scaffolds 63
4.3.1 Privileged Substructures 63
4.3.2 Target Community-Selective Scaffolds 64
4.3.3 Target-Selective Scaffolds 67
4.4 Target Promiscuity of Scaffolds 67
4.4.1 Promiscuous BM Scaffolds and CSKs 67
4.4.2 Scaffold–Target Family Profiles 70
4.4.3 Promiscuous Scaffolds in Drugs 70
4.5 Activity Cliff-Forming Scaffolds 71
4.5.1 Activity Cliff Concept 71
4.5.2 Multitarget Cliff-Forming Scaffolds 71
4.6 Scaffolds with Defined Activity Progression 73
4.6.1 Activity Profile Sequences 73
5 Exploring Virtual Scaffold Spaces 83
William R. Pitt and Boris Kroeplien
5.1 Introduction 83
5.1.1 Virtual Chemistry 83
5.1.2 Chemical Space 83
5.1.3 Scaffold Definition 84
5.2 The Comprehensive Enumeration of Parts of Chemical Space 85
5.2.1 Fragments 85
5.2.2 Ring Systems 86
5.2.3 Reagents 87
5.3 The Iterative Generation of Virtual Compounds 88
5.3.1 Transformations 88
5.3.2 Manual Selection of Chemical Modifications 88
5.3.3 Analog Generators 89
5.3.4 Inverse QSAR 89
5.3.5 Multiple Objective Optimization 90
5.3.6 Structure-Based De Novo Design 90
5.4 Virtual Synthesis 92
5.4.1 Synthetic Tractability 92
5.4.2 Using Real-Life Reactions for in Silico Molecule Construction 93
5.4.3 Readily Synthesizable Compounds 94
5.4.3.1 Construction 94
5.4.3.2 Searching 95
5.4.3.3 Outside Big Pharma 96
5.4.4 Iterative Approaches 96
5.5 Visualizations of Scaffold Space 96
5.6 A Perspective on the Past and the Future 97
References 99

Part Two Scaffold-Hopping Methods 105

6 Similarity-Based Scaffold Hopping Using 2D Fingerprints 107
Peter Willett
6.1 Fingerprints 107
6.2 Retrospective Studies of Scaffold Hopping Using 2D Fingerprints 109
### 6.3 Predictive Studies of Scaffold Hopping Using 2D Fingerprints

Page 112

### 6.4 Conclusions

Page 114

References Page 115

### 7 CATS for Scaffold Hopping in Medicinal Chemistry

Christian P. Koch, Michael Reutlinger, Nickolay Todoroff, Petra Schneider, and Gisbert Schneider

Page 119

#### 7.1 Chemically Advanced Template Search

Page 119

#### 7.2 Retrospective Evaluation of Enrichment and Scaffold Hopping Potential

Page 122

#### 7.3 Prospective Scaffold-Hopping Applications

Page 126

#### 7.4 Conclusions

Page 128

References Page 128

### 8 Reduced Graphs

Kristian Birchall

Page 131

#### 8.1 Introduction

Page 131

#### 8.2 Generating Reduced Graphs

Page 133

##### 8.2.1 Reduction Scheme

Page 133

##### 8.2.2 Node Labeling

Page 134

##### 8.2.3 Sheffield Implementations

Page 135

##### 8.2.4 Extended Reduced Graphs

Page 136

#### 8.3 Comparison and Usage of Reduced Graphs

Page 137

##### 8.3.1 Conventional Fingerprinting

Page 138

##### 8.3.2 RG-Specific Fingerprints

Page 139

##### 8.3.3 Augmenting Fingerprints with Edit Distance

Page 140

##### 8.3.4 Extended Reduced Graph Fingerprints

Page 141

##### 8.3.5 Graph Matching Approaches

Page 143

##### 8.3.6 Bioisostere Encoding

Page 144

#### 8.4 Summary

Page 146

References Page 146

### 9 Feature Trees

Nathan Brown

Page 149

#### 9.1 Introduction

Page 149

#### 9.2 Feature Tree Generation

Page 149

#### 9.3 Feature Tree Comparison

Page 150

#### 9.4 Retrospective Validation

Page 151

#### 9.5 Implementations and Applications

Page 152

##### 9.5.1 MTree: Combinations of Query Molecules

Page 152

##### 9.5.2 Similarity Searching in Large Combinatorial Chemistry Spaces

Page 152

#### 9.6 Conclusions

Page 153

References Page 154
10 Feature Point Pharmacophores (FEPOPS) 155
Jeremy L. Jenkins
10.1 Similarity Searching in Drug Discovery 155
10.2 FEPOPS: An Analogy to Image Compression 157
10.3 Computing FEPOPS 159
10.4 Scaling and Correlations 162
10.5 Defining Scaffold Hopping 163
10.6 FEPOPS in Similarity Searching and Scaffold Hopping 164
10.7 Alternative Alignment 168
10.8 In Silico Target Prediction 170
10.9 Chemical Space Uniqueness 171
10.10 Perspective on FEPOPS’ 10 Year Anniversary 172
References 173

11 Three-Dimensional Scaffold Replacement Methods 175
Nathan Brown
11.1 Introduction 175
11.2 Generic Three-Dimensional Scaffold Replacement Workflow 175
11.2.1 Molecule Databases 175
11.2.2 Fragment Generation and Filtering 177
11.2.3 Fragment Replacement Search and Scoring 178
11.3 SHOP: Scaffold HOPping by GRID-Based Similarity Searches 178
11.4 ReCore 179
11.5 BROOD 179
11.6 Conclusions 180
References 181

12 Spherical Harmonic Molecular Surfaces (ParaSurf and ParaFit) 183
David W. Ritchie and Violeta I. Pérez-Nuño
12.1 Introduction 183
12.2 Spherical Harmonic Surfaces 183
12.3 Rotating Spherical Polar Fourier Expansions 185
12.4 Spherical Harmonic Surface Shape Similarity 186
12.5 Calculating Consensus Shapes and Center Molecules 187
12.6 The ParaSurf and ParaFit Programs 188
12.7 Using Consensus Shapes to Probe the CCR5 Extracellular Pocket 190
12.8 Conclusions 192
References 193

13 The XED Force Field and Spark 195
Martin Slater and Andy Vinter
13.1 Pharmacological Similarity – More than Just Chemical Structure 195
13.2 Improving the Generation of Valid Molecular Fields 199
Contents

13.3 The eXtended Electron Distribution (XED) Force Field 200
13.4 The XED Force Field Applied to Scaffold Hopping in Spark 202
13.5 How Spark Works 202
13.6 Application of Spark in Drug Discovery Scenarios 206
13.7 P38 Kinase Inhibitor Fragment Growing Using Spark 207
13.7.1 The Beauty of P38 207
13.8 Creating New Molecules 208
13.9 New Potential Inhibitors 210
13.10 The Far-Reaching Consequences of Using Molecular Fields as Measures of Similarity 212
References 213

14 Molecular Interaction Fingerprints 215
Didier Rognan and Jérémy Desaphy
14.1 Introduction 215
14.2 Target-Annnotated Ligand Fingerprints 215
14.2.1 Interacting Atom/Fragment Fingerprints 216
14.2.2 Protein–Ligand Pharmacophores 217
14.3 Ligand-Annnotated Target Fingerprints 217
14.4 True Target–Ligand Fingerprints 220
14.4.1 Association Fingerprints 220
14.4.2 Interaction Pattern Fingerprints 222
14.5 Conclusions 225
References 226

15 SkelGen 231
Nikolay P. Todorov
15.1 Introduction 231
15.2 Structure Generation and Optimization 232
15.2.1 Fragments and Fragment Sets 232
15.2.2 Structure Generation 234
15.2.3 Scoring and Optimization 234
15.2.4 Ligand-Based Design 234
15.3 Validation Studies 235
15.3.1 Retrospective Validation Study: CDK2, COX2, ER, MMP-3 235
15.3.2 Estrogen Receptor 235
15.3.3 Histamine H3 Inverse Agonists 236
15.4 Scaffold Hopping Using Fixed Fragments 237
15.5 Scaffold Hopping Using Site Points 238
15.6 Further Considerations for Scaffold Hopping 240
15.6.1 Receptor Flexibility 240
15.6.2 Water Molecules 241
15.6.3 Receptor Specificity 242
15.7 Conclusion 242
References 243
Case Study 1: Scaffold Hopping for T-Type Calcium Channel and Glycine Transporter Type 1 Inhibitors

Leah C. Konkol, Timothy J. Senter, and Craig W. Lindsley

Introduction 247
T-Type Calcium Channel Inhibitors 248
Scaffold Hopping to Access Novel Calcium T-Type Channel Inhibitors 250
Scaffold Hopping to Access Novel Glycine Transporter Type 1 (GlyT1) Inhibitors 253
Conclusions 255

References 255

Case Study 2: Bioisosteric Replacements for the Neurokinin 1 Receptor (NK1R)

Francesca Perruccio

Introduction 259
Neurokinin 1 (NK1) Therapeutic Areas 259
The Neurokinin 1 Receptor (NK1R) and Its Mechanism 260
Neurokinin 1 Antagonists 261
NK1 Receptor: Target Active Site and Binding Mode 264
Bioisosteric Replacements in NK1 Receptor Antagonist 266
Bioisosteric Replacements in NK1 Receptor Antagonist: A Retrospective Study 270
Summary and Conclusions 274
References 274

Case Study 3: Fragment Hopping to Design Highly Potent and Selective Neuronal Nitric Oxide Synthase Inhibitors

Haitao Ji and Richard B. Silverman

Fragment-Based Drug Design 279
Minimal Pharmacophoric Elements and Fragment Hopping 281
Fragment Hopping to Design Novel Inhibitors for Neuronal Nitric Oxide Synthase 283
Fragment Hopping to Optimize Neuronal Nitric Oxide Synthase Inhibitors 288
Application of Neuronal Nitric Oxide Synthase Inhibitors to the Prevention of Cerebral Palsy 289
References 291

Index 297
List of Contributors

**Jürgen Bajorath**
Rheinische Friedrich-Wilhelms-Universität
Department of Life Science Informatics, B-IT
LIMES Program Unit Chemical Biology and Medicinal Chemistry
Dahmannstr. 2
53113 Bonn
Germany

**Kristian Birchall**
MRC Technology Centre for Therapeutics Discovery
1-3 Burtonhole Lane
London NW7 1AD
UK

**Julian Blagg**
The Institute of Cancer Research
Division of Cancer Therapeutics
Cancer Research UK Cancer Therapeutics Unit
15 Cotswold Road
Sutton, Surrey SM2 5NG
UK

**Nathan Brown**
The Institute of Cancer Research
Division of Cancer Therapeutics
Cancer Research UK Cancer Therapeutics Unit
15 Cotswold Road
Sutton, Surrey SM2 5NG
UK

**David Anthony Cosgrove**
AstraZeneca
Discovery Sciences
Chemistry Innovation Centre
Mereside 30S391
Alderley Park
Macclesfield SK10 4TG
UK

**Jérémy Desaphy**
UMR 7200 CNRS/Université de Strasbourg
Laboratoire d’Innovation Thérapeutique
MEDALIS Drug Discovery Center
74 route de Rhin
Illkirch 67400
France
List of Contributors

Ye Hu
Rheinische Friedrich-Wilhelms-Universität
Department of Life Science Informatics, B-IT
LIMES Program Unit Chemical Biology and Medicinal Chemistry
Dahlmannstr. 2
53113 Bonn
Germany

Jeremy L. Jenkins
Novartis Institutes for BioMedical Research
Developmental and Molecular Pathways
220 Massachusetts Avenue
Cambridge, MA 02139
USA

Haitao Ji
University of Utah
Department of Chemistry
Center for Cell and Genome Science
315 South 1400 East
Salt Lake City, UT 84112-0850
USA

Christian P. Koch
Eidgenössische Technische Hochschule (ETH)
Department of Chemistry and Applied Biosciences
Institute of Pharmaceutical Sciences
Wolfgang-Pauli-Str. 10
8093 Zürich
Switzerland

Leah C. Konkol
Vanderbilt University Medical Center
Department of Chemistry
2213, Garland Avenue
Nashville, TN 37232-6600
USA

Boris Kroepflien
UCB Celltech
208 Bath Road
Slough SL1 3WE
UK

Sarah R. Langdon
The Institute of Cancer Research
Division of Cancer Therapeutics
Cancer Research UK Cancer Therapeutics Unit
15 Cotswold Road
Sutton, Surrey SM2 5NG
UK

Craig W. Lindsley
Vanderbilt University Medical Center
Department of Chemistry
2213, Garland Avenue
Nashville, TN 37232-6600
USA

Violeta I. Pérez-Nueno
Harmonic Pharma
615 Rue du Jardin Botanique
54600 Villers-lès-Nancy
France

Francesca Perruccio
Francesca Perruccio
Novartis Pharma AG
Postfach
CH-4002 Basel
Switzerland

William R. Pitt
University of Cambridge
Department of Biochemistry
80 Tennis Court Road
Cambridge CB2 1GA
UK
List of Contributors

Michael Reutlinger
Eidgenössische Technische Hochschule (ETH)
Department of Chemistry and Applied Biosciences
Institute of Pharmaceutical Sciences
Wolfgang-Pauli-Str. 10
8093 Zürich
Switzerland

Petra Schneider
Eidgenössische Technische Hochschule (ETH)
Department of Chemistry and Applied Biosciences
Institute of Pharmaceutical Sciences
Wolfgang-Pauli-Str. 10
8093 Zürich
Switzerland

David W. Ritchie
Inria Nancy – Grand Est
Team Orpailleur
615 Rue du Jardin Botanique
54600 Villers-lès-Nancy
France

Timothy J. Senter
Vanderbilt University Medical Center
Department of Chemistry
2213, Garland Avenue
Nashville, TN 37232-6600
USA

Didier Rognan
UMR 7200 CNRS/Université de Strasbourg
Laboratoire d’Innovation Thérapeutique
MEDALIS Drug Discovery Center
74 route de Rhin
Illkirch 67400
France

Richard B. Silverman
Northwestern University
Department of Chemistry
Chemistry of Life Processes Institute
Center for Molecular Innovation and Drug Discovery
2145 Sheridan Road
Evanston, IL 60208-3113

Gisbert Schneider
Eidgenössische Technische Hochschule (ETH)
Department of Chemistry and Applied Biosciences
Institute of Pharmaceutical Sciences
Wolfgang-Pauli-Str. 10
8093 Zürich
Switzerland

Martin Slater
New Cambridge House
Bassingbourn Road
Litlington
Cambridgeshire
SG8 0SS
UK

Nickolay Todoroff
Eidgenössische Technische Hochschule (ETH)
Department of Chemistry and Applied Biosciences
Institute of Pharmaceutical Sciences
Wolfgang-Pauli-Str. 10
8093 Zürich
Switzerland
List of Contributors

Nikolay P. Todorov
Molscape Research Limited
145–157 St John Street
London EC1V 4PW
UK

Andy Vinter
New Cambridge House
Bassingbourn Road
Litlington
Cambridgeshire SG8 0SS
UK

Peter Willett
University of Sheffield
Information School
211 Portobello Street
Sheffield S1 4DP
UK
Preface

In 1999, Gisbert Schneider coined the term “scaffold hopping” for a systematic approach to modify the molecular skeleton of a lead structure [1]. Whereas in bioisosteric replacement atoms or small groups are substituted by other ones with identical or at least similar stereoelectronic features [2], scaffold hopping exchanges the central part of a molecule by a molecular frame of similar shape and pharmacophoric pattern [3]. Correspondingly, scaffold hopping may be considered as an extension of bioisosteric replacement. In this manner, it provides a conceptual and practical route for generating new chemistry and lead series with higher efficacy, better or modified selectivity, and/or improved pharmacokinetic properties, based on known active principles.

As often in science, this approach is also not completely new. The modification or exchange of a molecular scaffold was already applied in the chemical variation of morphine, quinine, some steroid hormones (e.g., estradiol), and β-blockers, to list only a few examples. Looking at naturally occurring β-lactams, that is, the penicillins, cephalosporins, and monobactams, we see that also nature sometimes uses this principle. Like in “fragment-based design,” where a breakthrough came only after the description of the advantages of this method, the definition “scaffold hopping” appealed medicinal chemists to use this strategy – and fueled its systematic application in lead structure search and optimization. Marketed analogs of celecoxib (Celebrex®), sildenafil (Viagra®), and several kinase inhibitors are recent examples of drugs and clinical candidates resulting from this approach.

The volume is logically organized in three parts. An introductory part deals with the representation, diversity, and navigation aspects of scaffold hopping. The next section is dedicated to topological methods, feature trees, shape-based methods, three-dimensional scaffold replacement methods as well as pharmacophore- and structure-based methods of scaffold hopping. Finally, some case studies demonstrate the value of scaffold hopping in all important target classes, exemplified by the design of ligands of the T-type calcium channel, the glycin transporter type 1, the neurokinin 1 receptor, and nitric oxide synthase.

The series editors highly appreciate that after editing the first monograph Bioisosteres in Medicinal Chemistry, Nathan Brown also undertook the effort to edit this monograph. We are very grateful that he organized this work, cooperating with so many excellent authors. Surely this book adds another fascinating new facet...
to our book series on “Methods and Principles in Medicinal Chemistry.” Last but not least, we thank Wiley-VCH, in particular Frank Weinreich and Heike Nöthe, for their valuable contributions to this project and the entire series.

Düsseldorf  
Weisenheim am Sand  
Zurich  
July 2013

Raimund Mannhold  
Hugo Kubinyi  
Gerd Folkers

References


A Personal Foreword

“…I want to stand as close to the edge as I can without going over. Out on the edge you see all the kinds of things you can’t see from the center.”

*Player Piano* (1952)
*Kurt Vonnegut, Jr.*

The foundation of a medicinal chemistry project is the determination and selection of the molecular scaffolds from which the potential drugs are grown. Therefore, it is essential that this fundamental core element be selected appropriately and with careful consideration. The selection of scaffolds and identification of ideal replacement scaffolds can be greatly assisted by computational analyses.

This book is the first to be dedicated to the analysis of molecular scaffolds in drug discovery and the discussion of the plethora of computational approaches that have been reported in scaffold hopping. Scaffold hopping is a subset of bioisosteric replacement where one tries to replace the core motif of a molecule while retaining important interaction potential, whether functionally or literally the necessary scaffolding for decorating with functional substituents.

There has been much published on what constitutes a molecular scaffold over the last century since the advent of Markush structures. A medicinal chemist tends to know it when they see it, whereas computational scientists apply various algorithms to identify what may be the scaffold of a chemical series or individual molecules. Part One of this book covers fully the many considerations in molecular scaffold identification, representation, diversity, and navigation. These are essential definitions and analyses that are prerequisites for application in scaffold hopping campaigns.

Part Two of this book, and the most substantial, covers a well-established subset of the many different computational methods that have been developed and applied in recent years. These range from ligand-based topological pharmacophores to abstracting three-dimensional structures in a variety of ways, including the use of protein structures.

Finally, and of key importance to the presentation of any approach, the book concludes with three chapters in Part Three in which scaffold hopping techniques and approaches have been applied prospectively in real projects. These case studies
consider scaffold hopping applied to designing ligands in four targets: nitric oxide synthase, the neurokinin 1 receptor, the T-type calcium channel, and the glycine transporter type 1.

I would like to extend my personal thanks to the contributors of all of the chapters in this book who have devoted so much time and effort in producing work that is of the high standard that we have come to expect in this book series “Methods and Principles in Medicinal Chemistry.” I would like to thank the series editors Raimund Mannhold, Hugo Kubinyi, and Gerd Folkers for commissioning to edit this book and also the previous book Bioisosteres in Medicinal Chemistry. Finally, I would like to thank the Wiley-VCH team for helping me pull this book together and making my life as editor a lot simpler in many ways; in particular, I would like to thank Frank Weinreich and Heike Nöthe for their invaluable efforts.

This book has been a labor of love for me and I am delighted that this book has formed so well through the duration of this project. I can only hope that you as the reader get as much out of reading it as I did in editing.

London, 2013

Nathan Brown
Part One
Scaffolds: Identification, Representation Diversity, and Navigation
1
Identifying and Representing Scaffolds

Nathan Brown

1.1
Introduction

Drug discovery and design is an inherently multiobjective optimization process. Many different properties require optimization to develop a drug that satisfies the key objectives of safety and efficacy. Scaffolds and scaffold hopping, the subject of this book, are an attempt to identify appropriate molecular scaffolds to replace those that have already been identified [1,2]. Scaffold hopping has also been referred to as lead hopping, leapfrogging, chemotype switching, and scaffold searching in the literature [3–6]. Scaffold hopping is an approach to modulating important properties that may contravene what makes a successful drug: safety and efficacy. Therefore, due consideration of alternative scaffolds should be considered throughout a drug discovery program, but it is perhaps more easily explored earlier in the process. Scaffold hopping is a subset of bioisosteric replacement that focuses explicitly on identifying and replacing appropriate central cores that function similarly in some properties while optimizing other properties. While bioisosteric replacement is not considered to a significant degree in this book, a sister volume has recently been published [7], many of the approaches discussed in this book are also applicable to bioisosteric replacement.

Some properties that can be modulated by judicious replacement of scaffolds are binding affinity, lipophilicity, polarity, toxicity, and issues around intellectual property rights. Binding affinity can sometimes be improved by introducing a more rigid scaffold. This is due to the conformation being preorganized for favorable interactions. One example of this was shown recently in a stearoyl-CoA desaturase inhibitor [8]. An increase in lipophilicity can lead to an increase in cellular permeability. The replacement of a benzimidazole scaffold with the more lipophilic indole moiety was recently presented as a scaffold replacement in an inhibitor targeting NS5B polymerase for the treatment against the hepatitis C virus [9]. Conversely, replacing a more lipophilic core with the one that is more polar can improve the solubility of a compound. The same two scaffolds as before were used, but this time the objective was to improve solubilility, so the indole was replaced for the benzimidazole [10]. Sometimes, the central core of a lead molecule can have
pathological conditions in toxicity that needs to be addressed to decrease the chances of attrition in drug development. One COX-2 inhibitor series consisted of a central scaffold of diarylimidazothiazole, which can be metabolized to thiophene S-oxide leading to toxic effects. However, this scaffold can be replaced with diarylthiazolotriazole to mitigate such concerns [11,12]. Finally, although not a property of the molecules under consideration per se, it is often important to move away from an identified scaffold that exhibits favorable properties due to the scaffold having already been patented. The definition of Markush structures will be discussed later in this chapter and more extensively in Chapter 2.

Given the different outcomes that lead to what can be called a scaffold hop, one can surmise that there must be different definitions of what constitutes a scaffold hop and indeed the definition of a scaffold itself. This chapter particularly focuses on identifying and representing scaffolds in drug discovery. Markush structures will be introduced as a representation of scaffolds for inclusion in patents to protect intellectual rights around a particular defined core, which will also be discussed in Chapter 2. Objective and invariant representations of scaffolds are essential for diversity analyses of scaffolds and understanding the scaffold coverage and diversity of our screening libraries. Some of the more popular objective and invariant scaffold identification methods will be introduced later in this chapter. The applications of these approaches will be discussed in more detail later in this book, with particular reference to the coverage of scaffolds in medicinal chemistry space.

1.2 History of Scaffold Representations

Probably the first description, which is still in common use today, is the Markush structure introduced by Eugene A. Markush from the Pharma-Chemical Corporation in a patent granted in 1924 [13]. Markush defined a generic structure in prose that allowed for his patent to cover an entire family of pyrazolone dye molecules:

I have discovered that the diazo compound of unsulphonated amidobenzol (aniline) or its homologues (such as toluidine, xylidine, etc.) in all their isomeric forms such as their ortho, meta and para compounds, or in their mixtures or halogen substitutes, may be coupled with halogen substituted pyrazolones (such as dichlor-sulpho-phenyl-carboxlic-acid pyrazolone) to produce dyes which are exceptionally fast to light, which will dye wool and silk from an acidulated bath.

More specifically, Markush’s claims were as follows:

1) The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.
2) The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

3) The process for the manufacture of dyes which comprises coupling dichlor-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

Interestingly, the careful reader will note that claims 1 and 2 in Markush’s patent are exactly the same. It is not known why this would have been the case, but it may be speculated that it was a simple clerical error with Markush originally intending to make a small change in the second claim as can be seen in the third claim. Therefore, Markush’s patent may not have been as extensive since it is possible one of his claims did not appear in the final patent.

Markush successfully defended his use of generic structure definitions at the US Supreme Court, defining a scaffold together with defined lists of substituents on that scaffold. Extending the chemistry space combinatorially from this simple schema can lead to many compounds being covered by a single patent. However, there remains a burden on the patent holders that although it may not be necessary to synthesize every exemplar from the enumerated set of compounds, each of the compounds must be synthetically feasible to someone skilled in the art. A patent may not be defendable if any of the compounds protected by a Markush claim cannot subsequently be synthesized.

An example of a possible Markush structure for the HSP90 inhibitor, NVP-AUY922 (Figure 1.1a) is given in Figure 1.1b. However, an example of a medicinal chemist may determine as the molecular scaffold is given in Figure 1.1c [14,15].

The Markush claim discussed above is clearly a mechanism for extending the protection of a single patent application to a multitude of related and defined compounds. The earliest reference to what we would now call a molecular scaffold definition that this author could identify was in 1969, in an article published in the *Journal of the American Chemical Society*, which provided the following definition [16]:

The ring system is highly rigid, and can act as a scaffold for placing functional groups in set geometric relationships to one another for systematic studies of transannular and multiple functional group effects on physical and chemical properties.

Clearly, this is a simple description of what constitutes a molecular scaffold and is readily understandable to a scientist active in medicinal chemistry and a specific example of a structural scaffold. However, its simple definition belies an inherent challenge in the identification of molecular scaffolds. Quite often, a medicinal chemist can identify what they would refer to as a molecular scaffold. This often involves identification of synthetic handles. The challenge here though is to
Identifying and Representing Scaffolds

Figure 1.1 The HSP90 inhibitor NVP-AUY922 depicted using different scaffold representations. (Reproduced from Ref. [20].)
understand how the scaffold has been determined, but this is a soft problem that is not capable of being reduced to an objective and invariant set of rules for scaffold identification. An expert medicinal chemist will bring to bear a wealth of knowledge from their particular research foci during their career and knowledge of synthetic routes: essentially, their intuition. Given a molecule, there are many ways of fragmenting that molecule that may render the key molecular scaffold of interest for the domain of applicability.

1.3 Functional versus Structural Molecular Scaffolds

Scaffolds can be divided roughly into two particular classes: functional and structural. A functional scaffold can be seen as a scaffold that contains the interacting elements with the target. Once defined, medicinal chemistry design strategies can concentrate on further improving potency while also optimizing selectivity and other properties, such as improving solubility. Conversely, a structural scaffold is one that literally provides the scaffolding of exit vectors in the appropriate geometries to allow key interacting moieties to be introduced to decorate the scaffold.

1.4 Objective and Invariant Scaffold Representations

It is important to be able define objective and invariant scaffold representations of molecules not only to permit rapid calculation of the scaffold representations but to also allow comparisons between the scaffolds of different molecules. Much research continues into objective and invariant scaffold representations, but here we summarize some of the methods that have seen significant utility. These scaffold representations use definitions of structural components of molecules: ring systems (Figure 1.1d), linkers (Figure 1.1e), side chains (Figure 1.1f), and the framework that is a connected set of ring systems and linkers (Figure 1.1g).

1.4.1 Molecular Frameworks

One of the first approaches to generating molecular scaffolds from individual molecules was the molecular framework (often referred to as Murcko frameworks) and graph framework representations [17]. Here, each molecule is treated independently, therefore, the method is objective and invariant.

The molecular framework is generated from an individual molecule by pruning all acyclic substructures that do not connect two cyclic systems (Figure 1.1h). The graph framework is a further abstraction in which the atom labels and bond orders are discarded to provide a simple abstraction of the general topology of the
molecule. The molecular (or Murcko) and graph framework representations of NVP-AUY-922 are given in Figure 1.1h and i, respectively.

This work was the first approach to classifying the crude shapes of molecules in terms of their cyclic frameworks. The inclusion of these topological representations and calculations of equivalences were suggested as being ripe for application to the de novo design problem. The study also highlighted the lack of scaffold diversity based on these representations in drug-like molecules and concluded that this would be an area of interest for medicinal chemists to understand which frameworks are underrepresented. The framework definitions were also applied to analyze the scaffold diversity in the Chemical Abstracts Service registry of 24,282 compounds at the time of publication in 2008 [18]. This application will be discussed more thoroughly in Chapter 3.

1.4.2 Scaffold Tree

Schuffenhauer et al. [19] defined the scaffold tree as a set of prioritization rules to systematically prune a given molecule. Starting from the molecular framework defined by Bemis and Murcko [17], rings are sequentially removed using the prioritization rules until only a single ring remains, the so-called level 0 scaffold. The prioritization rules defined for the scaffold tree are provided in Table 1.1.

By application of each of the prioritization rules defined by the scaffold tree method, each molecule in a data set is represented as a directed linear path of iteratively pruned fragments. The scaffold tree pruning strategy is data set independent: a given molecule will always result in the same result. However, the generation of the scaffold tree itself is a summary of a given data set. The pruning path of each molecule in a data set is analyzed and paths merged with one another to generate one or more scaffold trees. For a given data set, one scaffold tree will be the result if all of those molecules in the data set have the same common single

<table>
<thead>
<tr>
<th></th>
<th>The prioritization rules defined to prune ring systems in the generation of the scaffold tree.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Remove three-member heterocycles</td>
</tr>
<tr>
<td>2</td>
<td>Retain macrocycles of greater than 11 members</td>
</tr>
<tr>
<td>3</td>
<td>Remove rings first by longest acyclic linker</td>
</tr>
<tr>
<td>4</td>
<td>Retain spiro, nonlinear, fused and bridged rings</td>
</tr>
<tr>
<td>5</td>
<td>Retain bridged over spiro rings</td>
</tr>
<tr>
<td>6</td>
<td>Remove rings of size 3, 5, and 6 first</td>
</tr>
<tr>
<td>7</td>
<td>Fully aromatic rings should not be removed if remaining system is not aromatic</td>
</tr>
<tr>
<td>8</td>
<td>Remove rings with fewest heteroatoms first</td>
</tr>
<tr>
<td>9</td>
<td>If (8) is equal, use precedence relationship of N &gt; O &gt; S</td>
</tr>
<tr>
<td>10</td>
<td>Remove smaller rings first</td>
</tr>
<tr>
<td>11</td>
<td>Retain saturated rings</td>
</tr>
<tr>
<td>12</td>
<td>Remove rings with a heteroatom connected to a linker</td>
</tr>
<tr>
<td>13</td>
<td>Tiebreaking rule based on alphabetic ordering of a canonical SMILES representation</td>
</tr>
</tbody>
</table>
ring, the level 0 scaffold. With each additional level of the scaffold tree, the rings are included from each of the molecules in reverse order of the pruning process. Therefore, the level 1 scaffold will typically contain two ring systems (although this is not the case for monocyclic rings).

The advent of the scaffold tree method provided a simple, yet interpretable, hierarchical classification of data sets of molecules using an objective and invariant structural pruning strategy. The authors in their original work postulated a number of applications of the scaffold tree, including the analysis of structure–activity relationships (SAR), particularly in the context of high-throughput screening (HTS) campaigns. The scaffold tree from a pyruvate kinase assay of 602 active and 50,000 inactive molecules is given in Figure 1.2. Analysis of compound collections offered by commercial compound vendors or of the internal compound collection of an organization is an approach to investigating the structural diversity of these libraries, which may or may not be desirable depending on the purpose of those libraries.

In 2011, Langdon et al. [20] published a scaffold diversity analysis using the level 1 of the scaffold tree compared with molecular frameworks across the range of compound libraries, including those from vendors, internal fragment and lead-like screening files, exemplified medicinal chemistry from the literature and database of marketed drugs. This work is presented in further detail by the authors of this study in Chapter 3.

The scaffold tree algorithm has more recently been extended to generate Scaffold Networks by some of the original authors of the study [21]. As the name implies, Scaffold Networks generate a highly interconnected network of relationships between molecules and their entire enumerated sets of fragments.

1.5 Maximum Common Substructures

The calculation of the maximum common substructure (MCS) of a given congeneric series of molecules is formally not solvable in polynomial time, although it can be approximated in most cases for chemical structures and used effectively [22].

The challenge of using MCS algorithms on congeneric series can be overcome largely by introducing an iterative clustering, based on molecular similarity, followed by application of an MCS algorithm, which iterates until a termination condition is satisfied regarding the quality of the MCS at each stage. Nicolaou et al. [23] published the first implementation of an iterative approach to calculating the set of MCSs over a scaffold heterogeneous data set. This iterative approach allowed the generation of MCS groups from large sets of diverse molecules typically found in HTS libraries.

Clark and Labute [24] apply the scaffold tree approach by Schuffenhauer et al. for the detection, alignment, and assignment of scaffolds. Once the scaffold tree is generated, a score is generated for each fragment in the tree according to
the fraction of the remaining molecule set that contains that fragment, number of heavy atoms in the fragment, theoretical number of R groups, number of fragments selected in previous iterations, and the similarity of the fragment to each previously selected fragment. The method published addresses multiple scenarios of databases with varying degrees of scaffold homogeneity, including
homogeneous single scaffolds, misleading nonscaffolds, multiple similar common scaffolds, ambiguous common scaffolds, symmetrical common scaffolds, overly common scaffolds, and user-specified scaffolds.

While the method developed by Clark and Labute is not an MCS algorithm in principle, the results were demonstrated to be closer to the expectations of a medicinal chemist.

1.6 Privileged Scaffolds

A scaffold is deemed to be privileged if it appears many times across multiple targets [25]. Privileged scaffolds were first referred to in 1988 as “privileged structures” [26]. However, the significance of its privilege may not be as a result of commonality in terms of function. Depending on what is decorating an identified scaffold, the function of the resultant molecule with decoration may be significantly different. Take the example of piperazine, which may be monosubstituted or disubstituted, its scaffolding impact can be very different if it is a spiro center or not. It is important to understand the context of the scaffold in terms of biological target and also to realize that a particular scaffold may have been explored more deeply in one medicinal chemistry project than the other for various reasons.

1.7 Conclusions

This chapter has introduced a number of, but not exhaustive, published methods for scaffold identification. While it is typically intuitive for an expert medicinal chemist to be able to identify the scaffold of a given molecule, this may not be the same scaffold identified by other similar experts. However, for computational analysis, it is desirable to have an objective and invariant definition of a scaffold. The objective and invariant identification of the molecular scaffold from either an individual molecule or a set of congeneric molecules remains an unsolved problem. This is essentially due to soft issues surrounding scaffold definitions as discussed, but algorithms have been developed that can identify and appropriate scaffold representation in most cases.

This book is structured into three distinct parts. Part One covers different approaches to scaffold representations, analysis of scaffold diversity, and navigating the scaffold space. In this part, concepts discussed briefly here will be expanded upon with more consideration given to Markush structures, analysis of the scaffold diversity, and mining and hopping in these data. Finally, the part concludes with approaches to exploring virtual scaffold spaces that can be enumerated.

Part Two represents a selection of scaffold hopping algorithms and methods that represent a subset of the current state of the art. This part covers methods that
utilize topological representations of molecules, molecular shape, pharmacophores, and explicit information from protein–ligand cocrystal structures.

Part Three includes a selection of recent case studies from successful medicinal chemistry efforts from recent publications to demonstrate how these approaches can be used to move a medicinal chemistry project forward using scaffold hopping techniques.

Acknowledgments

N.B. is funded by Cancer Research UK Grant No. C309/A8274.

References


2
Markush Structures and Chemical Patents

David Anthony Cosgrove

2.1
Introduction

A patent is an agreement between an inventor and society whereby the inventor discloses enough information about an invention to enable someone skilled in the art to reproduce it, and in return society grants the inventor exclusive rights to the invention for a limited period of time, typically 20 years. The purpose is to stimulate innovation. The inventor invests resources in producing the invention and the period of exclusivity allows him/her the opportunity of recovering the costs of the invention and deriving profit from it. Society learns the mechanism of the invention, and other members may use this knowledge to improve upon and vary the invention.

Chemical patents are an essential part of the drug discovery process. Due to the cost of developing new medicines (a contentious matter, but which 10 years ago was put as high as $800 million [1]), it is important to protect a proposed new drug invention from copies for a period of time in an attempt to recoup those development costs. In that respect, it is no different from any other invention barring the fact that the initial costs are somewhat high.

A Markush structure forms a central part of most chemical patents, particularly those from the pharmaceutical industry. These structures are a general description of a set of related chemical compounds, normally based around a central, fixed core with various variable substituent groups. The substituents are commonly given as a mixture of named structures, structure diagrams, and generic groups of structures such as “alkyl” or “heteroaromatic” that could encompass many hundreds or thousands of potential substituents, depending on how they are defined within the patent. Time and storage requirements mean that, while in principle possible, it is not generally feasible to enumerate all structures encompassed by the typical Markush structure from a patent.

The Markush structure is named after Eugene Markush, a chemist who in 1924 was granted a patent [2] on a method of preparing pyrazolone dyes for which the starting materials were described in general terms. The US Patent Office initially rejected his submission on the grounds that it did not describe specific...
compounds, but he successfully pursued his case to the Supreme Court. In common parlance, a Markush structure is normally taken to imply a structure diagram such as the one shown in Figure 2.1, but Markush’s original patent has no such diagram.

Given the economic importance of Markush structures and the desirability, when developing new medicines, of avoiding operating within the scope of other workers’ patents, it is perhaps surprising that they have received relatively little attention in the cheminformatics literature and even less in recent times. There was a flurry of activity in the 1980s and 1990s, but since then things have been relatively quiet, and recent publications have been either reviews of the field, such as the recent one by Downs and Barnard [3], or means of analyzing the results from existing systems [4–6].

In the 1980s and early 1990s, a group from Sheffield University developed a system for the encoding and searching of Markush structures, publishing some 24 papers on the subject. The last in the series [7] is a summary of the work and contains references to the previous 23. The system described consisted of a language, GENsAL, for describing the Markush, and programs that allowed the inputting of a Markush described in GENsAL to generate an Extended Connectivity Record (ECTR), programs for producing bit screens from the ECTR file, and programs for searching with a query structure or partial structure to identify which Markush structures included the query. Contemporaneously, the Chemical Abstracts Service (CAS)1 was extending its chemical database technology to handle Markush structures [8,9], giving MARPAT, and a collaboration between Derwent Publications Ltd, Questel SA, and INPI (the French Patent Office) was modifying its DARC system to produce Markush DARC with a similar end in mind. The latter is now part of the Merged Markush Service (MMS) of Thomson Reuters.2 MARPAT and MMS are still the only commercially available databases and associated search systems. They are both closed systems, and the detailed workings of both have not been disclosed. However, both CAS and Derwent contributed

1) Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210, USA. http://www.cas.org, accessed April 17, 2012.

Figure 2.1 Markush structure diagram from WO2004/096129, a Merck patent application for inhibitors of AKT.
financially to the Sheffield project, so it is reasonable to assume that they have
taken similar approaches.

The MMS and MARPAT systems are of great utility when attempting to establish
whether one's compound is within the scope of another’s patent. Typically, a search
with a specific compound will result in a list of a few hundred patents for which the
Markush structures might include the compound. The services are of limited use
in the detailed analysis of a single patent. In addition to Markush structures,
chemical patents normally contain names and/or structures of specific chemicals
that exemplify the invention and, increasingly, data on the biological effects of the
compounds. When starting work in a biological area, the additional information
may be of value in understanding which properties of the disclosed compound are
important to the biological activity, information that can be transferred to novel
chemical series. Several companies such as CAS\(^1\) and GVK Biosciences\(^3\) provide
services to abstract structures and associated biological data from chemical patents
but there is little specialized software available to assist in analysis of these data.

In 2008, the Periscope project was started at AstraZeneca to address some of
these shortcomings. Specifically, the aim of the project was to enable the encoding
of Markush structures into a machine-searchable form such that a query
compound could be matched against the Markush. Output from the query should
include an assessment of whether the compound falls within or outside the extent
of the Markush as encoded, but more importantly how the Markush maps onto the
compound – which part of the query molecule matches the core and which part
matches the R groups. In the common case of recursive R group definitions, where
an R group is defined in terms of sub-R groups that may themselves comprise
subsub-R groups, the system should also map these subgroups onto the query
molecule. The mapping of the Markush onto the query is of value both when the
query matches and when the query does not match the Markush. It is often as
useful to know which bits of a molecule are outside of the Markush definition as to
know how another falls within it.

Potential uses of such a system, not all of which are currently exploited at
AstraZeneca, include the following:

- Encoding competitor Markush structures to ensure an in-house project is outside
  the scope of claimed intellectual property (a freedom-to-operate (FTO) inquiry).
- When preparing one’s own patent application, one could encode the Markush
  structure to ensure that it is chemically and logically consistent and that all
  example compounds fall within it.
- The Markush structure taken from a patent application or derived by other means
  such as inspection of a series of compounds can be used to decompose a set of
  compounds into R groups and a Free–Wilson analysis \(^{10}\) carried out.
- As a powerful substructure search system able to perform searches such as “a
  pyridine substituted at the 2- or 4-position by a chain of one to three methylenes

finished by a five- or six-membered heterocycle containing at least one nitrogen and zero or one sulfur atoms,” which would be beyond most other substructure search systems.

The Periscope system comprises three components: a language, the Markush Input Language (MIL); a program Menguin (Markush Encoding Graphical User Interface) to assist in encoding a Markush structure into an MIL file; and a search program i3am (is it in a Markush) to allow one or more query structures to be matched against the contents of an MIL file. The system has previously been described in some detail [11], with emphasis on the language for encoding the Markush structures. This chapter gives more detail of the search algorithm and exemplifies its use for scaffold hopping.

2.2 Encoding Markush Structures

Cosgrove et al. have given a full description of the MIL and its supporting information in Ref. [11]. The interested reader is referred to this, with only a brief summary being given here.

The MIL is an XML-based language. XML is a text-based format that was originally developed for the transfer of data across the Internet in a file that is both human- and machine-readable. It is a series of tagged records that can be nested, and is well suited to data where there is varying content for different records, such as in a Markush structure. The files can be somewhat verbose, which can make them tedious to generate by hand, but it does mean that a reader with only passing knowledge of the language structure can readily understand the contents.

A Markush structure in its most general case consists of a core structure, substituted with R groups that may in turn be substituted by further R groups in a recursive manner, in principle to infinite depth (which would denote a polymer). This structure is reflected in the MIL, where the key data field has the tag r_group. A typical MIL has multiple r_group records, each specifying an instance of an R group in the associated Markush structure. The substitution pattern of the R group is described by way of substituent records, which give the name of the substituent R group along with any information required to denote the position, number of occurrences, and so on. There are three types of R groups: Exact, Inexact, and Fused.

Each MIL file must have at least one r_group record of type Exact named Core, which may or may not have an exact correspondence with the core of the Markush as drawn in a patent (see Section 2.2.3). The Core R group must be Exact so that the search program has a definite place to start when attempting to match the Markush to a query molecule. In principle, the Core could also be Inexact, but the search would be much slower as it would have to start building fragments from each atom in turn, making the search space much larger. We have yet to see an example where that additional ability has been necessary.
2.2.1

The r_group Record

All r_group records must have an attribute type, which defines whether it is Exact, Inexact, or Fused, and name that identifies it in the Markush. The name need not be unique, since it is common for each R group in a Markush to have multiple definitions, such as “Z is alkyl, alkenyl, alkynyl, heterocyclyl, . . . ” Each Z would be encoded with its own r_group record, each of which can be of any type. In addition, there are records attach_by and attach_by_bond, which specify how the R group attaches to the R group of which it is a substituent (its parent group), text and remark, which give short and longer descriptions for documentation purposes, several records for homologous chains of the type —[C(R)](R)_, and substituent for specifying details of any substituents of the R group.

2.2.1.1 Exact R Groups

Exact R groups are those for which it is possible to specify a defined substructure, by either a SMARTS definition⁴) or an Accelrys RGFile.⁵) The point of attachment to the parent may be specified by the sequence number of the corresponding atom in the substructure definition. The attachment points of bidentate R groups such as ethylenedioxy moieties may be defined by way of a complicated SMARTS-based extension of the attach_by record attributes.

2.2.1.2 Inexact R Groups

These are more general definitions than Exact R groups, and in most Markush structures, particularly those from patents, they form the bulk of the R group records. They are for encoding such R groups as “alkyl,” “heteroaryl,” and so on, where there is no clear substructure, just a set of rules defining what is and is not allowed to be classed as a match to the group. Fields in the Inexact record include atom_count, bond_count, and smarts_count, whereby limits on the numbers of different element and bond types and more general substructural features may be set. These are needed because the definitions used for alkyl, for example, are generally different from patent to patent and may even differ in different sections of the same patent. For example, alkyl might be defined as one to three carbons in one patent, and one to six carbons in another. Some groups, such as alkyl, may be branched or unbranched, cyclic or acyclic, and rings may be aromatic or aliphatic, and the MIL allows for these distinctions also.

2.2.1.3 Fused R Groups

Fused R groups are convenient for specifying fused ring systems in terms of their subrings. In many cases, it is possible to specify the precise parameters for a fused ring system using just the definitions from an Inexact R group, but this is frequently

tedious and occasionally not the case. Take, for example, the situation where one has a carbocycle defined as aliphatic, with three different ring sizes: four to five atoms, six atoms, and more than six atoms. In the four- to five-atom case, one double bond is allowed, for six atoms, two are allowed, and for larger rings, three are allowed. This is essentially allowing maximum unsaturation in the ring without it becoming aromatic. We might also have a case where an aliphatic heterocycle is defined with the same ring sizes and bond counts, but allowing one or two nitrogen, oxygen, or sulfur atoms. The Markush might then define a ring where the carbocyclic ring and the heterocyclic ring are fused. This gives rise to nine different possibilities, with a four- to five-atom carbocycle fused to the four- to five-atom heterocycle, and so on. With care, it would be possible to define this ring system from scratch using the Inexact R group syntax, although the restriction that if two heteroatoms are found, they must be in the same subring would not be possible without some extremely careful SMARTS definitions and would be unlikely to cover all eventualities. All nine possibilities would have to be coded individually. The Fused R group definition was developed to address this difficulty. It has two records, fused_group_one and fused_group_two, which are used to specify which ring R groups provide the two subring definitions, either of which can be a Fused R group allowing multiple fused ring systems to be defined. It is also possible to specify whether the fused system must attach to the parent by the first or second ring type, or if either is acceptable.

The Fused R group allows considerable flexibility and convenience when writing the MIL. However, as described in Section 2.3.1.3, this comes at the expense of increased complexity in the search.

2.2.2 The Menguin Program

It is possible to encode a Markush into the Markush Input Language by typing the requisite XML into a file using a simple text editor. However, the long-winded nature of the XML syntax makes it tedious and with the tedium comes errors. To address this, the program Menguin was developed. It allows the user to “drag and drop” R groups from a pre-prepared dictionary, editing them as required, and possibly developing a dictionary particular to the Markush in question. Where the dictionary does not contain the required definitions, they can be either sketched using Accelrys’ JDraw applet, in the case of Exact R groups or defined using a dialog sheet for Inexact ones. Further details are given in Ref. [11].

A typical patent generally takes less than a day to encode, the bulk of which is spent analyzing the text of the Markush and attempting to resolve ambiguities and errors. What should one do, for example, with statements like “The term ‘heteroaryl’ as used herein means a monocyclic heteroaryl or a bicyclic heteroaryl. The monocyclic heteroaryl is a five- or six-membered ring. The five-membered ring includes two double bonds and one, two, three, or four nitrogen atoms and

---

6) JDraw, v1.1.200.121, 2010. Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.
optionally one oxygen or sulfur atom . . . Representative examples of monocyclic heteroaryl include, but are not limited to, furanyl, . . . [12],” where furanyl seems not to meet the general description that appears to require at least one nitrogen atom? As another example, Markush’s original patent [2] claims “anilines and homologs of anilines.” This may have been clear to a chemist “skilled in the art” in 1924, but in 2012 requires further investigation to divine the scope of the claim. Elsewhere in the patent, Markush talks of xylidine and toluidine, so when encoding the patent, one might settle on aniline optionally substituted by one or more methyl groups. Because of situations like this, it appears unlikely that it will be possible in the near future to have a fully automated system for encoding a Markush structure from the text in a patent.

2.2.3 Correspondence between the MIL File and the Markush Structure

When encoding a Markush structure, the most important consideration should be ensuring that the MIL file captures the logical extent of the Markush, such that a trained chemist reading the description of the Markush and the search program i3am would place the same query structures within and without the Markush. In understanding the results from i3am, it is convenient if the structure of the MIL file follows the Markush as described in the patent, but this is not always possible. For example, in Figure 2.2, taken from patent application WO2006/021881 for protein kinase inhibitors, there are two areas where it is either not convenient or not possible to encode the Markush exactly as it is drawn. The group Y in the core can be either N or CR12. There is syntax in the MIL to allow this to be specified as an in-ring variation, but it is slightly cumbersome to use and in this instance it is easier simply to draw the two cores in JDraw within Menguin. If there were more variations in the ring, or more options for Y such that the number of different cores was much larger, it could well be easier to use the more flexible MIL constructs. However, when parsing the MIL, all that i3am does is build internally all the different core substructures, so the effect is much the same. If the core was such that in the variation where Y was C, the C was not substituted, one could also draw the substructure with an optional atom at that point, allowing the values C or N. In SMARTS notation, this would be the SMARTS string “[NH2]c1ncc[c,n]c1O”. This is perfectly valid in the MIL, but in this case one has lost the distinction between the two different core types. If one wanted to perform a Free–Wilson analysis, for example, to probe the effect of the two different cores on activity, this would not be

![Figure 2.2](image_url) Figure 2.2 Markush structure from patent application WO2006/021881 of Pfizer, Inc.
a useful way of performing the encoding. However, if the purpose of the encoding
were merely to screen a set of molecules to see whether or not they fell within the
Markush, this would be a convenient shortcut that would approximately halve the
run time for the searches.

The other area in this Markush where a direct correspondence between the
drawing and the MIL encoding is not possible is in the variable-length chain linking
A to the core, where there are between one and four methylene groups, each bearing
substituents R₉ and R₁₀. As a general rule, these linking groups must be dealt with as
a substituent on the core that bears another substituent A. In principle, the MIL and
i3am could use the same syntax as for the in-ring variation to describe this part of the
core. However, generally these sorts of linker groups are rather more complicated
than the one or two atom variations that are typical for rings, and dealing with all the
possibilities would add extra complexity to an already complex language. In this case,
one would encode the core as the anilinopyridine ring plus the ether oxygen, and
define a new substituent for the linking chain, substituted by A. By careful use of the
text and remark fields for these r_group records, the correspondence to the
Markush structure as drawn can be described such that it should be clear to future
users of the encoded MIL file exactly what has been done.

The situation is even more complicated in this case because the linking group is
also a homologous chain of repeating methylene units. The MIL has additional
constructs for specifying this type of substituent, allowing the encoder to specify
the number of repeating units in the chain, if appropriate which atoms are used to
connect the units together (to allow distinction between, for example, straight-chain
and branched propylene linkers) and whether substituents are on each unit or only
the end one (cf. R₉ and A). However, in this simple case it might be easier to adopt
the same philosophy as described earlier for Y, and simply draw the four
possibilities as different versions of the same R group definition.

2.3
The Search Algorithm

At its simplest, the search algorithm needs merely to be able to tell whether a query
compound is contained within the Markush structure. This is certainly sufficient
for large-scale screening of a database of Markush structures, such as may be found
in a database of patents. For more general uses, as described earlier, a more
informative output is required. Generally, one would wish to know which parts
of the query compound match which parts of the Markush, and, in the cases where
the query does not match the Markush, which parts are missing or surplus. For
example, if the Markush calls for two or three substituents on a group, and the
query contains only one, then it has a missing part. If it contained four substituents
at that position, one would be surplus. Both would be screening misses. For Free–Wilson analyses, it is convenient to be able to process multiple molecules in one
pass and output a table giving for each molecule, in some form, a description of the
parts of the molecule that matched each R group in the Markush.
The search program i3am is written in C++ using the OEChem cheminformatics toolkit\(^7\) for molecule manipulation, substructure matching, and so on. The search is a depth-first recursive algorithm and proceeds as shown in Algorithm 2.1. The operation of the algorithm is perhaps best understood by example.

**Algorithm 2.1: Outline of the i3am search algorithm**

**initialisation:** Set the R Group to be considered to ‘Core’ and the molecule portion under consideration to the whole molecule

**forall the R Groups at current level do**

**forall the Definitions of R Group do**

/* An R Group may have more than one definition, such as alkyl, alkoxy, halogen etc. */

Match R Group Definition against all remaining portions of molecule

**forall the Matches do**

| Subtract from the matched atoms from the molecule portion |
| Increment the score for this branch by the number of matched atoms |

if R Group has substituent R Groups and still portions of molecule remaining then

| Recurse to line 1 |
| end |

else

| return best score at this level |

end

Increment score with best score from lower level

if Current score > best score then

| Best score ← current score |
| Note matched atoms; Note matched R Group instance; |
| end |

if Best score = Number of atoms in molecule portion being matched then

| return best score |

end

end

end

if Number of R Groups < Number of portions then

| Note portions unmatched to R Group |
| Flag match as incomplete |
| end |

if Number of portions < Number of R Groups then

| Note R Groups unmatched to portion |
| Flag match as incomplete |

end

Find best scoring combination of R Groups mapping to molecule portions

| return score for combination |

end

---

\(^7\) OEChem, v1.7.4, 2012. OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508, USA.
Consider the case of the Markush structure shown in Figure 2.3. It has been chosen to have an unsymmetrical Core, with two R groups with a small number of options, one option for R₁ being further substituted by R₃. Molecule 1 in Figure 2.4 is matched against this Markush as follows:

1) The indole core is matched, and there is only one possible mapping.
2) The atoms of the indole core are removed from the molecule, leaving two pieces: an isobutyl and an ethoxy group.
3) Xenon atoms are added to the cleaved groups to show the point of attachment to the parent molecule fragment. This is so that subsequent levels in the search do not produce erroneous matches due to incorrect attachments. For example, if an amide is being matched, the algorithm should not match reverse amide fragments. Xenon was chosen because a real element was required by the OEChem toolkit, and xenon is rarely seen in organic chemistry. Equally, an unusual or impossible isotope of a different element could have been used as a tag.
4) There are still R groups and portions of molecule unmatched, so the algorithm recurses to the next level.
5) Because the positions of R₁ and R₂ are unambiguous in this case, R₁ is matched against the isobutyl moiety and R₂ is compared with the ethoxy group.
6) Match the different definitions of R₁ against the isobutyl piece. The first does not match, but the second does, so return score for that match.
7) Likewise for the ethoxy and the different R₂ definitions.
8) The two fragments match their respective R group definitions with no atoms left over, so the search is complete. The R1 group receives a score of 4 and R2 has a score of 3, which are returned to the previous level. These scores are added to the score for the Core, giving a total score of 16.

9) All possibilities at the Core level of the search have been considered, so the search is complete with Molecule 1 matching the Markush exactly.

For Molecule 2 in Figure 2.4, the search proceeds identically to step 5. In this case, removal of the Core leaves methoxycarbonyl and chloride groups. When R1 is matched to the methoxycarbonyl, the carbonyl group matches the first option for R1, leaving a methoxy group. Another level of search is required to see whether the methoxy group matches any of the possibilities for R3. It matches C1–alkoxy, with a score of 2; this is returned to the R1 level of the search, where it is added to the 2 from the carbonyl match, so that this level returns a score of 4. The total score for the molecule is then $9 + (2 + 2) + 1$, and again a complete match.

For Molecule 3 in Figure 2.5, the search returns an incomplete match at line 6 because the substituted carbonyl does not match at all, and removing the matches to the alkyl group leaves an extant chloride group. The latter match is the higher scoring one, so the partial match to the alkyl is recorded, and the chlorine atom in the full molecule is flagged as unmatched. For Molecule 4, there is nothing to match R1, so again the query returns a partial match, with R1 flagged as unsatisfied.

2.3.1

Matching R Groups

The search algorithm revolves around the matching of R group definitions to pieces of molecule. The procedure is different for the three types of R groups.

2.3.1.1 Exact R Groups

Exact R groups are the easiest to match, since the definition includes a substructure query in either SMARTS or RGFile format. The program i3am is implemented

![Figure 2.5](image-url) Further molecules matched against the Markush in Figure 2.3, which do not satisfy the definition.
using the OEChem toolkit, which has methods for returning all subsets of atoms that match the query. Some substructures will produce multiple possible matches for some molecules, which can form two types. The substructure might be present in two different places in a molecule, and there might be multiple ways of matching the substructure to the same atoms in a molecule. For example, a phenyl substructure matches the molecule biphenyl in 24 different ways. There are 12 ways in which the phenyl substructure can match each benzene ring in biphenyl (starting from each of the six atoms, and going either way round the ring), and there are two such rings. Each of these possible matches must be investigated in turn. Normally, however, the attachment point and substitution positions can be used rapidly to reduce the number of options passed to the next level of the search tree.

For example, in Figure 2.6, the 12 possible matches of the phenyl group (SMARTS c1ccccc1) are reduced to 2 that need to be investigated: one where R2 matches the methyl group and R3 matches the cyano group, and the other where

```
<name>R1</name>
<type>Exact</type>
<smarts>c1ccccc1</smarts>
<attach_by>1</attach_by>

<substituent>
  <subst_name>R2</subst_name>
  <subst_attach_point>3</subst_attach_point>
  <type>plain</type>
</substituent>

<substituent>
  <subst_name>R3</subst_name>
  <subst_attach_point>5</subst_attach_point>
  <type>plain</type>
</substituent>
```

Figure 2.6 Example of multiple matches of Exact R group to molecule portion. The Markush diagram for the R group is shown, along with the corresponding MIL text and a potential match. The xenon atom marks the point of attachment to the parent moiety.
the matches are reversed. Both will be scored, and the higher scoring mapping retained.

2.3.1.2 Inexact R Groups

The situation for Inexact R groups is more complicated, since there is no substructure, merely a set of rules that the molecular fragment must satisfy, such as atom and bond type counts, whether it is cyclic or not, and so on. Generally, many more possible matching fragments must be investigated. The procedure adopted in i3am is to generate all contiguous fragments of molecule from the point of attachment out from the parent moiety. This is done in a breadth-first recursive procedure, where each atom in the current fragment is examined in turn and all atoms that are directly connected to it and are neither in the parent nor yet in the fragment are added to a new copy. Thus, if an atom has three connected atoms, three more fragments are generated by incorporating each of these atoms into the existing fragment. Each fragment is then checked to see that it does not exceed the maximum limits for features defined in the R group definition. If it does, it is discarded and that branch of the fragment is not explored further. The fragment must be kept for further consideration if it does not yet meet minimum limits as the algorithm cannot look ahead to see which further atoms might be added to the fragment. Atoms are also only added to make larger fragments if they meet general features defined for the R group, such as aromaticity. For example, if the R group is defined as aromatic, an aliphatic atom or the one not in a ring will not be added.

For large substituents and general R group definitions, this part of the search can easily become the most expensive part. A possible optimization might be to generate just the largest fragments that meet the R group definition at each level in the search tree. However, this would not be correct because allowance must be made for the possibility of further substituents “needing” the same atoms. It is quite common for Markush structures from patents, which generally comprise long lists of possibilities for R groups, each substituted by another long list of possibilities, to contain such overlapping groups. For example, the Markush might have “R₁ is aromatic, heteroaromatic, cyclo alkyl, alkyl, . . . , substituted by R₅, where R₅ can be cyanide, alkyl, . . . , each such R₅ group being optionally substituted by halogen.” When matching this definition against Molecule 5 in Figure 2.7, if the search algorithm consumed all atoms matching the alkyl

---

Figure 2.7 Examples of overlapping R groups.
definition when dealing with R₁ before checking the remainder against R₅, it will find that the fragment does not match, since there is nothing that matches a definition for R₅. This incorrect fragmentation is shown in Molecule 6 in Figure 2.7. In fact, there are several ways in which the fragment can be divided so that there is an alkyl substituted by another alkyl substituted by halogen, such as the one shown in Molecule 7 in Figure 2.7, and the algorithm where all possible subfragments are generated will find them. As implemented in i3am, the search terminates when the first full match is found, but it could equally return all possibilities, at the expense of further computation time.

The problem of overlapping R group definitions is a significant one in the searching of Markush structures from chemical patents. The Sheffield project developed a sophisticated, multilayer set of fragment screens to enable the rapid searching of large databases of Markush structures, and a large part of the effort was devoted to addressing this issue [13–16], both for screen generation and for subsequent matching of query structures to the Markush. Their solution differs from that adopted here, as their system was geared more for the rapid screening of structures against a database of Markush structures rather than the detailed analysis of small numbers of query structures against a small number of Markush descriptions.

2.3.1.3 Fused R Groups

Fused R groups are composite R groups created by combining two different ring R group definitions, either of which might itself be a Fused R group. The matching of a Fused R group to a molecule fragment is carried out in two stages, using the relevant Exact or Inexact R group matching procedures as detailed earlier. All combinations of definitions for the two subgroups are examined. For those cases where the individual sets of atoms match the subgroups, further checking is required. The two subsets of matched atoms must form a contiguous fragment, with at least two atoms in common, the atoms at the fusion point. The atoms that are in common must be joined together by a continuous path that only incorporates other matched atoms. The attach_by property of the Fused R group specifies whether the group must attach by the first sub-R group or the second, or either. The two subsets of matched atoms must also be checked for this, as well as for the attach_by property of the subgroup itself. For example, if the Fused R group is to be attached by the first subgroup, and the first subgroup must be attached by a nitrogen atom, then both these must be satisfied for the molecule fragment to be deemed a match.

There is a particular subtlety in atom matching that must be accounted for when dealing with fused rings. Consider the situation where a Fused R group is defined as being a benzene ring fused to a nonaromatic heterocycle of five to seven atoms with one to three nitrogen atoms, which should match Molecule 8 in Figure 2.8. However, of the five atoms that match the nonaromatic heterocycle, two are in fact aromatic by virtue of their presence in the phenyl ring to which they are fused. The matching of the atom properties for the aliphatic ring must therefore be relaxed to allow for this, while maintaining the strict aliphatic properties for the other atoms in the ring.
2.3.1.4 Hydrogen Atoms

The importance of hydrogen bonds in molecular recognition events such as protein–ligand binding is well established [17]. One could argue that hydrogen is the most important element in ligand–receptor binding. Despite this, OEChem, in common with other cheminformatics toolkits, treats hydrogen atoms as an adjunct property of the heavy atoms to which they are attached. Generally, unless otherwise stated, the number of hydrogens attached to an atom is calculated from the normal valence of the atom and the number and orders of the bonds incident upon it. So, for the SMILES string [18] C, representing methane, four hydrogen atoms are implied, and for CC, ethane, three hydrogens are assumed attached to each carbon atom. Sometimes it is necessary to state explicitly the hydrogen atoms, so as to define chirality or resolve potential ambiguities. The SMILES for pyrrole, for example, is c1c[nH]cc1, because it would not be possible to determine from normal valence rules that the hydrogen atom is on the nitrogen. A similar approach to hydrogen atoms is taken in chemical patents. Generally, if no substituent is specified at a particular position, it is assumed that hydrogen atoms are present to satisfy the valence. Sometimes, however, hydrogens are explicitly included in the list of possible substituents required (i.e., nonoptional) at a given position. One means of dealing with this (adopted by the Sheffield system) is to treat such substituents as optional, even when in the text of the patent they were compulsory. While this is convenient and will not change the meaning of the encoded Markush, it means that a molecule deemed to be a match to the Markush will not necessarily appear so, as an R group described by the patent as compulsory will not be recorded as having a match in the query molecule.

In the Periscope project, the decision was taken to include these explicit hydrogens in the encoding and search results. This, however, presented problems with both the match search and results display. In the OEDepict module of the OpenEye toolkits, it is possible to display all hydrogen atoms, none, or some specific special cases such as hydrogen atoms required to define chirality. It is not possible to display a small number of arbitrary hydrogens. Equally, when matching atoms in a query structure against the R group definitions, it is not possible to label an arbitrary hydrogen atom as being matched. Indeed, the hydrogen atoms are not explicitly present in the connection table.

The approach taken in Periscope is to “promote” implicit hydrogen atoms to helium for the purposes of search and display. An explicit reference to hydrogen as a substituent is encoded in the MIL as an Exact R group for a helium atom with one
attachment point. This can be done by either SMARTS or RGFile record. When the search algorithm encounters such an R group in the Markush record, it “promotes” the requisite number of implicit hydrogen atoms to helium and adds them to the connection table. The helium atoms that are not matched in the search are subsequently removed and the implicit hydrogen count for the parent atom is reinstated. Helium was chosen because it is a safe assumption that it will not otherwise appear in either patents or chemical structures likely to be of interest to medicinal chemists. Therefore, if it appears in a search result, there will be no ambiguity about its origin. Also, when drawn in OEDepict, it appears as the atomic symbol He, which can be interpreted as “Hydrogen (explicit).”

2.3.1.5 Managing Multiple Fragment/R Group Matches
As the search proceeds, it is frequently the case that there are multiple molecule fragments to be matched against multiple R group definitions. For example, in Figure 2.2 there are two R groups attached to the core. When the search algorithm removes the core from a query structure, there are likely to be fragments of molecules left that must be matched to the R1 and R2 definitions. In Figure 2.2, the positions of the substituents are stated explicitly, so it is unambiguous which fragment must be matched to which R group definition. In Figure 2.1, there are between zero and two R2 groups at variable positions on the phenyl ring, and they are less clear. In this more general case, where there may be substitution at optional positions, it is necessary to establish the highest scoring assignment of matching fragments to R groups.

To this end, all fragments are matched against all R group definitions and a score is given for each match. If the position of the R group is given explicitly, then fragments that are not attached to the correct atom in the parent moiety are given a score of 0 without performing the match. If not, the full recursive match is carried out. A table is assembled where the rows are the R groups and the columns are the molecular fragments, and a cell contains the score of matching that fragment to that R group. If multiple occurrences of an R group are allowed, such as in the case “R1 is optionally substituted by one to three R5 groups,” a separate row is created for each possibility. The table so produced may have more columns than rows, in which case the result will be a partial match since at least one fragment will not be matched to an R group, or more rows than columns, which may or may not indicate a full match, depending on the optional nature of the R groups.

Finding the optimum combination of R groups and fragments such that the score is maximized is an example of a linear assignment problem (LAP). The original LAP was described for the situation where one has m computations to perform, each with an estimate of time required, and n computers, each with a relative speed. The problem is to find the allocation of computation to computer, which gives the least elapsed time for completion of all of them. In i3am, we use the algorithm of Carpaneto et al. [19], which has previously been used in other areas of cheminformatics [20,21], to solve the assignment. This is an $O(n^3)$ algorithm, but in i3am $n$ is rarely as high as 4. Indeed, in most cases it is 1 or 2, in which case the LAP algorithm is somewhat of an overkill.
When scoring the fragments for their match against the R group, each atom that matches is given a score of 100, with the exception of a matched explicit hydrogen, which will have been turned into a helium atom as described earlier and which receives a score of 99. The reason for this is demonstrated by the simple Markush structure shown in Figure 2.9. Molecule 9 is the result of a match where R₁ is hydrogen, with the matched hydrogen being shown by the helium atom. Molecule 10 is the result when R₁ is hydroxyl. The first step when the search algorithm reaches a comparison with R₁ is to remove an implicit hydrogen atom from the terminal methyl group and add a helium instead. In the case of Molecule 10, this gives a hydroxyl and a helium on the terminal carbon atom in the side chain. The search algorithm can then test the R₁ against the helium atom, which matches the first definition, followed by the hydroxyl group that matches the second definition, or vice versa. The order in a particular case is determined by the input order of the atoms in the structure: different orderings of the atoms will result in the helium being tested first or the hydroxyl. If both matches give the same score, it will appear to the user somewhat arbitrary which result is returned. However, the results are not equivalent. In the case where the hydroxyl matches, the unmatched helium lapses back to an implicit hydrogen, and the whole molecule is deemed to have matched the Markush. In the case where the helium matches, the hydroxyl is left unmatched, and the molecule is incorrectly deemed not to match the Markush. By giving the helium atom a slightly lower score, this problem is averted, but in cases such as Molecule 9, where it is the only possible match, it is still recorded.

2.4 Using Periscope for Scaffold Hopping

In 2004, Böhm et al. described the aim of scaffold hopping as being “to discover structurally novel compounds from known active compounds by modifying the
central core structure of the molecule.” The Periscope system can assist in this matter in three related ways:

1) The MIL can be viewed as a highly flexible and specific substructure search specification language that can be used to encode a core substructure in a general way in an attempt to find structurally similar replacement cores.

2) The Markush structure from a patent along with the example structures and activities can be used to derive a Free–Wilson analysis that will give insight into the roles of different parts of the molecules in the biological activity. This can assist in identifying what the core is, which substituents are essential for activity, which substituents might have been added to improve physical properties, and so on. A similar approach can be used with an arbitrary series of compounds with activities, although in this case the Markush structure will have to be defined at the start of the process. With a patent, it has already been done.

3) Given one or more patents for compounds active against a particular biological target, Periscope can be used to ensure that compounds in a project remain clear of the chemical space of any granted patents, as a precursor to a full freedom-to-operate search.

2.4.1 Substructure Searching

As an example, consider the Bextra patent WO1996/025405, with the Markush shown in Figure 2.10. Suppose one wished to find new substructures that retained the oxazole ring, but replaced the two phenyl rings with five- or six-membered aromatic heterocycles. This is very difficult to encode as a normal substructure query, where, while it is trivial to allow for atoms in the aromatic rings to be a heteroatom or carbon, it is not possible to specify at least one heteroatom in each ring. It can, however, be encoded as a simple MIL query. Searching a snapshot of PubChem taken on September 3, 2010, some 25 million compounds took about 5 h on a single processor of a reasonably modern Linux machine and produced 65 hits, some of which are shown in Figure 2.11. None of the hits contained the required
sulfonamide group, but they did contain a variety of different ring structures, which is the main requirement for such a scaffold-hopping exercise.

2.4.2 Free–Wilson Analysis

A Free–Wilson analysis [10] is a type of quantitative structure–activity relationship (QSAR) in which a multilinear regression model is built linking the presence or absence of different fragments in a molecule to a quantitative property (e.g., biological activity or solubility) of the molecule. It is applicable to any series of analogs where there is a large common substructure with limited variation, and where it can be assumed that the effects due to the fragments at different positions in the molecule are independent of each other. Patel et al. [22] have shown that this latter assumption is not always justified. Nevertheless, Free–Wilson analysis is often used as a way of gaining insight into the roles different parts of the molecules in a compound series play in their activities.

Figure 2.11 Hits from MIL substructure search of PubChem.
Frequently, the Free–Wilson analysis is performed by identifying a core substructure and using the substituents of the different molecules as they attach to the core as the descriptors for the regression equation. However, it need not be this limited, and, indeed, in their original paper, Free and Wilson themselves also allowed for a variable linker group $R_4$, either nothing or an amide linker, between the core and one of the substituents, as shown in Figure 2.12. Breaking down larger substituents into smaller pieces, while more time-consuming if performed manually, can sometimes improve the quality of models. If the core is small and the substituents are large, it is quite likely that individual substituents at a position will appear only in a small number of molecules, running the risk of overfitting of the model. By breaking the substituent down, this is less likely, with the added bonus that the model becomes more informative about what exactly is the effect of the different parts of the substituent.

In addition to a Markush structure, it is almost always the case that a chemical patent gives one or more specific structures that exemplify the series, and it is increasingly the case that numerical values for biological activity in an assay are also included. Together, these contain the basis of a Free–Wilson analysis: the Markush structure can be used rapidly to decompose the example structures into a table of fragments that can be combined with the given activity to produce a QSAR.

As an example, the Markush structure depicted in Figure 2.13 and associated exemplary structures from patent application WO03/072557, abstracted by GVK Biosciences, will be used. This patent has rather sparse activity data, so for the purposes of this demonstration, calculated log octanol/water partition coefficients (ClogP) have been used, as calculated by the ClogP program of

![Figure 2.12](image1.png)  
**Figure 2.12** Series of indanamines, as used in Table III of Ref. [10].

![Figure 2.13](image2.png)  
**Figure 2.13** Markush structure from Novartis patent application WO03/072557.
Clearly, a model that calculates a value calculated by another model is of limited practical value, and it is offered here only as an example of how the analysis might proceed. The patent was encoded into an MIL file, a process that took approximately 2h. The 180 example structures, represented as SMILES strings, were analyzed using i3am and a table giving the SMILES string for the matching portions of each molecule was produced. Only 179 of the structures were found to match the Markush, as encoded, with example 9 being deemed a non-match. This does not appear to be an error with the encoding. The multilinear regression model was built using the “Fit Model” procedure in JMP v.10. A plot of predicted versus actual ClogP from the model is shown in Figure 2.14, along with a piece of the prediction equation. Both the $R^2$ (0.97) and the root-mean-square error (RMSE) (0.2243) show a good, predictive model for these molecules. It is more than likely that it is overfitted; no attempt has been made to address this or optimize the model. The equation for the predictions allows one to draw conclusions about the effect the different groups have on ClogP. For example, absence of a group at R1 reduces the ClogP by 0.38 units, while adding propan-1-one increases it by 0.43 units. An interesting homologous series effect is also shown, where progressing from the methylamide substituent to the propylamide increases the predicted ClogP by 0.81 units. Were this an equation linking biological activity to the Markush structure, this could well be valuable data to transfer into other chemical series as part of a scaffold-hopping exercise.

9) JMP, v8, 2008. SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, USA.
2.4.3

Fast Followers

A so-called fast-follower project is one where one uses data from one or more published patents or patent applications to kick-start a drug discovery project of one’s own. The patent or patents provide information on one or more classes of compounds known to be active against a particular disease target from which valuable structure–activity relationships may be derived. In this sense, the patent applications are acting as an adjunct information source to the more conventional published literature. If there are multiple patent applications in an area, there is richness of information, but this can also present problems. It is desirable to remain outside the scope of other granted patents, so as to maintain novelty, but it can be difficult to keep in mind all the claims in a congested area. If the Markush structures from the relevant patent applications were encoded in searchable form, it would be relatively straightforward to check any proposed compounds against these encodings before commencing extensive experimental work.

Stewart et al. have described a system Drug Guru [23] that consists of a set of transformations commonly encountered in drug discovery and a method of applying them to molecules of interest. One can envisage a system whereby this was combined with an encoding of one or more Markush structures such that one or more compounds, for example, the examples from the patent application whence the Markush structure came, are transformed in an automated, iterative way using something akin to the Drug Guru program until structures were produced that fell outside Markush structures as deduced using i3am. This could lead to an efficient way of scaffold hopping within a fast-follower project in order to gain a novel chemical series.

2.5

Conclusions

Despite the importance of Markush structures in the preparation of chemical patent applications, research into their encoding and searching has not been widely reported. Generally, the matter has been left in the hands of a small number of commercial organizations that provide an indexing and searching service. The Periscope system described herein and elsewhere is the first significant work reported in this area for some years. Given sufficient investment, it would be possible to provide a large-scale service of the type offered by CAS or Thomson Reuters. However, as described earlier, there is also value to be had in the careful encoding and analysis of a small number of selected patents of particular interest to individual drug discovery programs, and this is the direction in which development of the Periscope system has initially been directed.

In addition to the practical value of having encoded patents available for, for example, Free–Wilson analysis, the encoding process can itself be of assistance in understanding the claims in a chemical patent. Having to decide exactly which
groups to include when encoding the Markush, or which elements and bond types are included or excluded in the definition of a particular group, means that one truly has to read the patent properly with close attention to detail, which can increase the amount of learning extracted from the patent. To this end, it is of value if the encoder is directly involved in the project, either by encoding it directly himself or being in close contact with the encoder rather than having the encoding performed as a service by a third party.

References


3
Scaffold Diversity in Medicinal Chemistry Space

Sarah R. Langdon, Julian Blagg, and Nathan Brown

3.1
Introduction

The aim of scaffold hopping is to replace the central core of a bioactive molecule with a structurally dissimilar fragment while retaining the activity of the molecule [1–3]. A major application of scaffold hopping is to access previously unexplored regions of chemistry space [1–3]. Scaffold hops into new territories of chemical space may be desirable to avoid potential issues in intellectual property rights and to explore new scaffolds that are potentially more interesting (e.g., improved synthetic accessibility or absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties), yet fulfill the same functional role [1–3].

Common approaches to scaffold hopping include analog-by-catalog and virtual screening techniques [1–3]. Both these approaches require a library of available compounds, usually an in-house screening collection, or a library of compounds that are commercially available. The extent to which chemical space can be explored via these methods depends on the diversity of these libraries. In order to identify previously uncharacterized scaffolds, the compound libraries being mined need to represent a large and diverse area of chemical space.

This chapter discusses the concept of scaffold diversity in medicinal chemistry relevant libraries. This is a difficult concept to examine; it is dependent upon how we define a scaffold and how we define diversity. We review analyses of the scaffold diversity of medicinal chemistry space as exemplified in screening, fragment, vendor, and drug libraries. These analyses explore the effect of different scaffold representations and definitions of diversity on the apparent scaffold diversity of medicinal chemistry space.

3.1.1
Scaffold Representation

To assess the scaffold diversity of a library, we must first identify the scaffolds of the compounds in the library. There are many ways of identifying the scaffold; these methods can be person dependent (e.g., Markush structures Figure 3.1b) [4], data
Figure 3.1 Examples of scaffold representations of the kinase inhibitor lapatinib. (a) Lapatinib, (b) Markush [7], (c) ring systems, (d) maximum common substructure, (e) molecular framework, (f) hetero framework, (g) graph framework, and (h) scaffold tree.
set dependent (e.g., maximum common substructure Figure 3.1d) [5], or independent of both the person defining the scaffold and the data set the compound occurs in (e.g., molecular frameworks) [6]. The optimal representation is often problem dependent.

When analyzing the scaffold diversity of a library, the scaffold representation can make the library appear more or less diverse. For example, a small generic scaffold could result in a library appearing to contain only a few scaffolds, with many examples of each scaffold, whereas a more granular scaffold representation could result in the same library appearing to contain many scaffolds, with fewer examples of each scaffold. For this reason, it is important to understand the scaffold representation used to assess scaffold diversity.

3.1.2 What Do We Mean by Scaffold Diversity?

Scaffold diversity can be quantified using a variety of methods, for example, the number of unique scaffolds in a library, the distribution of compounds over unique scaffolds in a library, or the structural diversity of unique scaffolds in a library. These methods may lead to different conclusions on scaffold diversity for the same library. For example, consider two libraries containing the same number of compounds, libraries A and B, Figure 3.2. If libraries A and B both contain the same number of scaffolds, all of which are equally represented, we would consider the libraries to be equally diverse. However, if we measure the structural similarity of the scaffolds, library A could contain structurally related scaffolds and library B could contain structurally dissimilar scaffolds, leading to the conclusion that library A is less diverse than library B (Figure 3.2a). If, on the other hand, library A contains more unique scaffolds than library B, it may be considered more scaffold diverse. However, library B may contain the same number of examples of each scaffold, whereas library A may contain many examples of one scaffold and only a few examples of the remaining scaffolds. Thus, although library A contains more scaffolds than B, one scaffold is heavily represented, therefore it could be considered less diverse than library B (Figure 3.2b). This example illustrates how the definition of scaffold diversity can affect the interpretation of scaffold diversity of a library; most studies discussed in this chapter use a combination of methods in an attempt to give a balanced view of scaffold diversity.

3.2 Scaffold Composition of Medicinal Chemistry Space

One method of assessing the scaffold diversity of a compound library is to determine the scaffold composition of the library; this involves identifying the scaffolds present in the library and the frequency of occurrence of these scaffolds. A common method for defining scaffolds for such an analysis is the molecular framework (Figure 3.1e). In 1998, Bemis and Murcko published their analysis of
5120 commercially available drugs [6]. Bemis and Murcko dissected compounds into ring systems, linkers, and side chains; the framework of a molecule is the union of the ring systems and linkers. A molecular framework retains information on atom and bond type (Figure 3.1e), whereas in a graph framework, all atoms are reduced to carbon and all bonds are considered as single bonds (Figure 3.1g). For the set of 5120 drugs, 1179 and 2506 unique graph and molecular frameworks were identified, respectively [6]. Of the 2506 molecular frameworks identified, 42 occur in 10 or more compounds and represented a total of 1235 compounds, 24% of the total data set. Also, 76% (1908) of the molecular frameworks occur in only one compound. Similar uneven representation of graph frameworks is also observed. These results show that the library of drugs contains a small number of well-represented scaffolds and many singleton scaffolds that only represent one compound. These observed characteristics of the drug data set have also been observed in many other studies of medicinal chemistry relevant data sets using

Figure 3.2 Diagram showing scaffold representation of libraries A and B. The enclosed box represents scaffold space and circles represent scaffolds. The closer the scaffolds are in scaffold space, more structurally similar the scaffolds are. (a) Libraries A and B both contain 100 molecules that are represented by 10 equally populated scaffolds. Library A contains scaffolds that are structurally similar, whereas library B contains scaffolds that are more structurally diverse. (b) Libraries A and B both contain 100 molecules. Library A contains 11 scaffolds, one scaffold represents 90 molecules, and the remaining 10 scaffolds represent one molecule each. Library B contains five scaffolds that represent 20 molecules each.
frameworks as well as other scaffold representations, a selection of which is discussed below.

Lipkus et al. conducted a similar analysis based on the scaffold composition of the CAS (Chemical Abstracts Service) registry [8]. The CAS registry is a database of chemical substances found in the scientific literature, which contained nearly 26 million substances when the analysis was performed in 2007; the analysis was performed on a set of 24,282,284 molecules that were classified as organic compounds. The analysis identified molecular frameworks and graph frameworks but focused on hetero frameworks that retain the same atom information as molecular framework but no information on bond type (Figure 3.1f). Lipkus et al. identified 836,708 graph frameworks, 2,594,334 hetero frameworks, and 3,380,334 molecular frameworks. This exemplifies how the scaffold representation can affect the outcome of such a diversity analysis; a more descriptive scaffold representation (i.e., molecular frameworks) will identify more unique scaffolds than a less-descriptive representation (i.e., graph frameworks). The analysis went on to examine the frequency of occurrence of the hetero frameworks. The most common 5% of the hetero frameworks represented 75.5% of the compounds, again illustrating that the library under consideration contains some very well-represented and many underrepresented scaffolds. For example, it was found that 12.7% of the organic compounds studied were represented by the 10 most frequently populated hetero frameworks.

These two seminal studies focus on a small set of drug compounds and a large set of organic compounds not necessarily representative of medicinal chemistry space. Constraints applied while designing screening libraries may remove many of the compounds considered in the CAS registry analysis and the set of approved drugs is only a very small subset of the explored medicinal chemistry space. Some studies have extended their analysis from approved drugs to drug candidates, clinical trial compounds, and bioactive molecules to include more of the potential medicinal chemistry space.

Xu and Johnson used molecular frameworks encoded with molecular equivalence indices [9] to analyze the scaffold composition of a set of 1,490 drugs from DrugBank [10] and 250,000 bioactive compounds selected from PubChem [11]. It was found that the drug set contained a higher proportion of unique frameworks (39.7%) than the bioactive set (29.4%), suggesting that drug compounds are more diverse than bioactive compounds. Despite these observations, 58.0% of the drug frameworks were singletons, while 48.1% of bioactive frameworks were singletons. This suggests that more unique scaffolds are present in the drug set. This could be attributed to the fact that only a small number of bioactive compounds evolve to become approved drugs, reducing the number of examples of some scaffolds in drug compounds. Thus, both sets contain a high proportion of singleton scaffolds and, consistent with the findings of other analyses, there are a few highly populated scaffolds. Similar observations were noted for a combined compound library of

approved drugs and drug-like compounds [12] and for 16 libraries of bioactive compounds against different biological targets [13].

Two studies of drugs and drug-like compounds examined the occurrence of fragments; in these analyses, a molecule may contain more than one fragment, unlike the molecular framework studies where a molecule contains only one framework [14,15]. Once again, some fragments occur more frequently in compound libraries. The analysis of the public database of the National Cancer Institute2) shows that particular combinations of fragments are more common than others [14]. These so-called “chemical clichés” were proposed as representatives of well-explored medicinal chemistry space.

Medicinal chemistry space can also be defined as the screening libraries used for high-throughput screens or as the vendor libraries from which screening compounds are purchased. Similar scaffold composition analyses have been performed on such libraries and have returned results consistent with the studies discussed above.

A scaffold composition study based on MCS was carried out on 17 screening libraries taken from 12 different suppliers, giving a total of 2.4 million compounds [16]. The MCSs identified in the libraries were categorized as classes if they represented at least two compounds, or as singletons if they represented only one compound. It was found that for all libraries there were more “singletons” than “classes” and that the distribution of molecules over “classes” was highly skewed with a few very highly populated scaffolds. Several metrics are used to assess the distribution of compounds over scaffold in the libraries; these will be discussed in detail in Section 3.3.

A recent scaffold diversity analysis compared seven medicinal chemistry libraries including a vendor library, small molecule and fragment screening libraries, the ChEMBL database,3) and DrugBank [10] data sets [17]. This study used both molecular frameworks and level 1 scaffolds from the scaffold tree [18] (Figure 3.1h) to assess scaffold composition. This is a key example of how the scaffold representation has an effect on apparent scaffold diversity. Both representations once again showed that the libraries consisted of a few highly populated scaffolds and many singleton scaffolds but this observation was more prominent when using the level 1 scaffolds. The level 1 scaffold is a less-granular representation than the molecular framework. Therefore, more unique molecular frameworks appear in the data set than level 1 scaffolds and the data sets appear more scaffold diverse when using a molecular framework definition.

All these studies present the same conclusion that medicinal chemistry relevant libraries are composed of a few highly represented scaffolds and many singleton scaffolds. Several theories have been put forward to account for why the same areas of scaffold space are more heavily explored than others. One suggestion is that biological activity is limited to a small area of chemical space [19]. In this case, it is

3) ChEMBL Database, https://www.ebi.ac.uk/chembl/.
proposed that more knowledge will be acquired for biologically active scaffolds thus making them more attractive for future use, whereas less-common scaffolds will continue to be less-popular options [14]. However, a recent study of publicly available bioactivity data has identified scaffolds that have selectivity for target gene families, but are underrepresented in approved drugs [20]; perhaps these scaffolds are underused due to the paucity of accumulated knowledge on them. Another theory is that some scaffolds are more synthetically accessible and/or commercially available than others, making them more attractive for lead development; in addition, some scaffolds may be preferred due to the relative synthetic ease of attaching additional groups in multiple substitution vectors [14,19]. Finally, it should be noted that the explored medicinal chemistry space is a small fraction of the total chemistry space, such that the scaffolds explored in medicinal chemistry cover only a small proportion of the available scaffold space [19].

The scaffold composition of medicinal chemistry space has been assessed by comparing libraries of drugs and bioactive compounds with screening and vendor libraries. Shelat and Guy studied the overlap in scaffold composition of drug, bioactive, and commercially available screening libraries to assess whether or not screening libraries have biological relevance [21]. Using molecular frameworks as the scaffold representation, it was found that libraries of commercially available bioactive molecules have a large scaffold overlap with drug scaffolds. In contrast, screening libraries made up of compounds that comply with Lipinski’s rule of five [22] or diversity-oriented synthesis compounds cover only a small proportion of drug scaffolds. Thus, drugs and bioactive compounds may have similar scaffold compositions as they are active against the same, well-explored target families; however, rule of five and diversity-oriented synthesis libraries may have use against hitherto unexplored biological targets [21]; such libraries may not represent the known medicinal chemistry space, but may be a source of novel chemotypes relevant for new therapeutic target classes.

3.2.1 Natural Products as a Source of Novel Medicinal Chemistry Scaffolds

As discussed above, medicinal chemistry relevant libraries contain scaffolds that have been heavily explored and are highly populated. It may be advantageous to introduce less-explored scaffolds into medicinal chemistry screening libraries to find structurally different bioactive compounds and to discover hit matter for underexplored biological targets. Natural products could be a source of novel scaffolds not represented in the known medicinal chemistry space.

Several efforts have been made to assess the scaffold composition of natural product libraries and compare these to the scaffold composition of drugs and small-molecule libraries. Lee and Schneider identified ring systems (Figure 3.1c) present in a collection of 10 495 natural products and 5757 marketed drugs [23]. It was found that 35% of drug-derived ring systems were also found in the natural product ring systems whilst only 17% of natural product-derived ring systems were also found in drug rings. These results illustrate that natural products
contain scaffolds previously unseen in medicinal chemistry space that may be starting points for novel hit matter.

Grabowski et al. performed a molecular framework scaffold composition analysis similar to that discussed above [12] where a natural product library was included alongside a drug and drug-like library. The library of 128,600 natural products contained 31,050 molecular frameworks (4.14 molecules per scaffold) and the library of 121,593 drugs/drug-like compounds contained 51,337 molecular frameworks (2.37 molecules per scaffold). Thus, the drug/drug-like compounds contain more unique scaffolds and are more diverse than the natural products. A high proportion of frameworks in both libraries are singletons, 59.4 and 69.2% for the natural product and drug/drug-like libraries, respectively. This slightly higher proportion of singletons in the drug/drug-like library could explain why a higher number of unique frameworks were observed for this library. These observations are also consistent with Singh et al. who conducted a molecular framework comparison of drug and natural product libraries [11]. Although the Grabowski study showed that the natural product library was less diverse than the drug/drug-like library, 90% of the natural product molecular frameworks are unique to the natural product library. This is consistent with the finding of Lee and Schneider [23], again suggesting that natural products may be a source of novel scaffolds to increase diversity in medicinal chemistry libraries.

3.2.2
Enumerating Potential Medicinal Chemistry Scaffolds

The scaffold composition of medicinal chemistry libraries has also been analyzed by enumerating all potential scaffolds or scaffold topologies and assessing how well the existing libraries cover all the enumerated possibilities. Apart from assessing the diversity of medicinal chemistry space, such a method can identify “holes” that are not represented in medicinal chemistry library space. Although a useful method, one drawback is that it is difficult to fully enumerate all possible scaffolds in medicinal chemistry space, and some studies enumerate only a particular type of scaffold such that the virtual library may not represent all the possible molecules.

Lipkus used a topological descriptor space, which describes the connectivity of atoms within a molecule, to assess the diversity of chemical ring systems in the CAS registry [24]. Three integer descriptors were used to characterize the aspects of ring topology; these three descriptors can be used as three dimensions of a topological-descriptor space. A nonredundant library of 474,859 ring systems found in the CAS registry as at the end of 1998 was plotted onto this topological-descriptor space that contained a total of 7387 cells. The distribution of ring systems over the possible topologies is uneven. Ninety percent of cells had a population less than 10; however 60 cells had a population of over 100 and contained almost 24% of all ring systems, once again illustrating that compound libraries are often distributed over a small range of common scaffolds, with many rarer or singleton scaffolds present. One interesting observation in this study is that the highly populated topologies are spread across the topological space, leaving significant voids. The topologies of
these voids have not been well explored in organic chemistry and could prove to be sources of novel ring systems of interest in medicinal chemistry; however, such voids may result from synthetic inaccessibility.

Wester et al. performed a similar analysis in which they enumerated the possible scaffold topologies and examined how well several medicinal chemistry libraries represent these topologies [25]. The scaffold topologies were created in an exhaustive enumeration of ring systems containing up to eight rings; these rings could be directly linked via fused, bridged, or spiro atoms, or they could be linked via linear linkers. This gave a total of 1,547,689 topologies [25]. A merged data set of seven libraries containing over 25 million compounds was found to contain only 23,737 of the possible topologies; half of the molecules in each library were represented by only eight topologies and between 62.8 and 91.3% of the libraries are represented by 18 topologies. This again illustrates that the chemistry space of the libraries is distributed over a few common topologies, although these topology analyses account only for the graph connectivity of scaffolds and exclude the information on atom type and bond order. Therefore, these libraries will reduce to a smaller number of topologies than more descriptive scaffolds such as molecular frameworks. As with Lipkus’ topology analysis, the work from Wester et al. identifies many gaps in topological space. However, it was observed that the merged library covers all possible topologies that contain 1–3 rings, the majority of possible topologies that contain four and five rings (98.6 and 85.9%, respectively), and 31.0% of topologies that contain six rings. For topologies containing seven and eight rings, the coverage drops below 4%. This illustrates that larger, and more complex, topologies are underrepresented whereas topologies of interest to medicinal chemists are well represented. Again, high synthetic complexity, as well as the high probability of poor drug-like properties, is likely to influence the paucity of compounds representing topologies containing >5 rings.

The work discussed above focuses on simple topological representations of scaffolds; similar studies have been carried out that retain information on atom and bond type. Ertl et al. enumerated a set of simple aromatic ring systems that contain one or two fused rings giving a total of 580,165 scaffolds; only 780 of these scaffolds were present in a set of 150,000 bioactive compounds [19]. The scaffolds were clustered and arranged into a self-organizing map in order to visualize the scaffold space. The 780 bioactive scaffolds were distributed sparsely over the scaffold space into well-defined activity islands. In a similar study, Pitt et al. generated the VEHICLe database, a set of 24,847 enumerated heteroaromatic ring systems [26]. Only 1701 of these ring systems were identified as synthesized, leaving a large area of possible scaffold space unexplored. Thus, the inclusion of atom and bond types in the topology analysis decreases the observed coverage of possible scaffold space for simple aromatic scaffolds containing one or two fused rings. We have previously stated that enumerated scaffolds may not be represented in the known chemistry space because they are synthetically inaccessible, but Pitt and colleagues used a machine-learning approach to predict that 3000 unrepresented ring systems could easily be synthesized. Both of these studies suggest that the known chemistry space occupies only a very small proportion of the potential chemistry space, and
that there is a wide area of potential chemistry space yet to be explored in medicinal chemistry. However, these two studies focus on simple aromatic ring systems; the libraries studied may contain compounds that contain more complex or saturated scaffolds that have not been represented in this work. Thus, although the medicinal chemistry space lacked many of the enumerated scaffolds, it may contain many more scaffolds that were not represented in the potential scaffold space generated in these studies.

3.2.3 Using Scaffold Composition to Interpret Bioactivity Data

We have discussed the use of scaffold representations to understand the scaffold composition and diversity of medicinal chemistry space. Scaffold representations have also been used to extract structure–activity relationships (SARs) from bioactivity databases.

Schuffenhauer et al. clustered the NCI cancer data set and in-house Novartis screening data using several different scaffold representations: molecular frameworks, the scaffold tree, and their scaffold identification and naming system (SCINS) code [26,27]. It was found that these scaffold representations had the ability to group molecules into clusters with similar biological profiles. The hierarchical nature of the scaffold tree also allowed it to be used to abstract more detailed SAR information, allowing activity changes to be attributed to structural changes that can be visualized within the hierarchical representation of compound libraries [28,29].

Bajorath and coworkers have done extensive work exploring scaffold–activity relationships [30]. Their work has included the identification of scaffolds that represent compounds with high binding affinity and selectivity for a particular target or target family [30–33]. Such scaffolds are often defined as privileged substructures [34]. Promiscuous scaffolds lie on the opposite end of the selectivity scale and have also been investigated by Bajorath et al. who identified scaffolds that represent compounds having biological activity against multiple target families [30,33,35]. Activity landscapes represent SAR by combining information on structure and activity of compounds [36]. Activity landscapes have been used to identify activity cliffs, where structurally similar compounds have a large difference in potency [36]. Bajorath et al. have analyzed activity landscapes and cliffs in the context of scaffolds to identify scaffold–activity relationships [37–42]. Chapter 4 discusses these interesting scaffold–activity relationships in more detail.

3.3 Metrics for Quantifying the Scaffold Diversity of Medicinal Chemistry Space

We have examined how the number of scaffolds in a library and their frequency can provide an indication of the scaffold diversity of medicinal chemistry space. This section looks at methods for quantifying scaffold diversity.
The studies discussed so far have identified scaffolds in the data set and counted the number of representatives of each scaffold. These studies can be extended further by looking at the distribution of compounds over these scaffolds.

Cumulative scaffold frequency plots (CSFP) are visual representations of the distribution of molecules over scaffolds. The scaffold frequency is defined as the number of molecules that contain a particular scaffold; this can be represented as a percentage of the total number of compounds in the data set. Scaffolds can then be sorted by their scaffold frequency, from most frequent to least frequent. A CSFP is generated by plotting the cumulative percentage of molecules against the cumulative scaffold frequency also represented as a percentage. In an evenly distributed data set where each scaffold is represented by the same number of molecules, the CSFP will be diagonal from (0%, 0%) to (100%, 100%) (Figure 3.3). Therefore, the closer the plot is to the diagonal, the more evenly distributed the data set is.

CSFPs have been used in many of the analyses discussed above and these plots further support the conclusion that medicinal chemistry space is unevenly distributed with a few well-represented scaffolds and many singleton scaffolds [8,13,16,17]. An exemplar CSFP is shown in Figure 3.3. This plot is typical of those presented in the studies that have used this technique. The curve begins with a very

![Figure 3.3](image-url) CSFP plot for the ChEMBL database represented with molecular frameworks (blue) and level 1 scaffolds (green).
steep gradient – the most heavily populated scaffolds that represent a large proportion of the scaffold library. The curves then become significantly shallower reaching a continuous gradient that represents the large number of singleton scaffolds present in the library. These plots can be used to compare the distribution of different libraries, or the same library with different scaffolds. Figure 3.3 shows a CSFP for the ChEMBL database represented with both molecular frameworks and level 1 scaffolds. The molecular framework curve is less steep than the level 1 scaffold curve, and levels out at a higher percentage of scaffolds. This shows that the data set is more evenly distributed over molecular frameworks than level 1 scaffolds. A proposed reason for the difference in scaffold distribution between the two scaffold representations is that molecular frameworks are a more granular scaffold definition and therefore there are more unique molecular frameworks in the data set than level 1 scaffolds (Table 3.1) [17]. Another way to compare the CSFPs of the two data sets is to calculate the area under the curves (AUC) [13]. An evenly distributed library with a CFSP curve on the diagonal between (0%, 0%) and (100%, 100%) will have an AUC of 0.5; less-evenly distributed libraries will have an AUC greater than 0.5, with a value of 1.0 representing the case where all molecules are represented by one scaffold. For the ChEMBL database represented with molecular frameworks and level 1 scaffolds, these values are 0.83 and 0.93, respectively (Table 3.1).

CSFPs can also be used to further quantify the distribution of molecules or scaffolds in a library; they can be used to calculate the percentage of classes (in this case, classes are defined as scaffolds) that represent $n$ percent of compounds ($PC_nC$); for example, $PC_{50}C$ is the percentage of scaffolds that represent 50% of compounds. $PC_{50}C$ was originally introduced by Krier et al. in their scaffold diversity analysis of screening libraries and is a helpful metric to compare the distribution of several libraries [16]. This concept has also been extended to give the percentage of scaffolds that represent 25 and 75% of the library to give $PC_{25}C$ and $PC_{75}C$, respectively [17].

Table 3.1 gives $PC_nC$ values for the ChEMBL database when represented with molecular frameworks and level 1 scaffolds. The $PC_nC$ values quantify the uneven distribution of the compound library over scaffolds. For both scaffold representations, the values are very low, showing that a large number of molecules are represented by only a small proportion of scaffolds. For example, the $PC_{75}C$ for level 1 scaffolds is 2.90; this means that 75% of molecules in the library are represented by only 2.90% of scaffolds. The $PC_nC$ values for the level 1 scaffolds are

| Table 3.1 Scaffold diversity analysis of the ChEMBL database using molecular frameworks and level 1 scaffolds. |
|-------------------------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Molecules                                      | Scaffolds                    | $PC_{25}C$                   | $PC_{50}C$                   | $PC_{75}C$                   | AUC                          |
|------------------------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Molecular frameworks                           | 866 943                      | 217 653                      | 0.49                         | 4.97                         | 23.04                        | 0.93                         |
| Level 1 scaffolds                              | 866 682                      | 99 387                       | 0.034                        | 0.30                         | 2.90                         | 0.83                         |

$PC_nC$ is the percentage of scaffolds that represent $n$ percent of molecules. AUC is the area under the curve of the corresponding CSFP plot (Figure 3.3).
lower than the corresponding values for molecular frameworks (2.90 versus 23.04% for PC75C), again illustrating how molecular frameworks give a more evenly distributed representation of the library than level 1 scaffolds.

Another metric for quantifying the scaffold diversity of a library is the Shannon entropy (SE); the Shannon entropy is an information-based entropy measure that can be used to quantify the distribution of compounds over scaffolds \[13,43,44\]. The Shannon entropy is defined as

\[
SE = -\sum_{i=1}^{n} p_i \log_2 p_i; p_i = c_i/P,
\]

where \(p_i\) is the relative frequency of the scaffold \(i\) in a population of \(P\) compounds containing a total of \(n\) unique scaffolds; \(c_i\) is the scaffold frequency of a scaffold. \(SE\) can take a value between zero and \(\log_2 n\); therefore, \(SE\) can be normalized between zero and one to give scaled \(SE\) (SSE) by dividing \(SE\) by \(\log_2 n\). An SSE value of zero corresponds to the case where all compounds are represented by one cyclic system; an SSE value of one corresponds to the case where all compounds are uniformly distributed over scaffolds. Medina-Franco et al. \[13\] used SSE to characterize the distribution of compounds over the 5, 10, and 20 most frequent scaffolds in 16 libraries of compounds active against different biological targets.

Kelley et al. developed a function to measure the spread of clusters of protein structures with related conformations and a function to measure the distance between different clusters \[45\]. These functions can also be used to measure the spread and distance of clusters of compounds \[26\], and give insight into the diversity of the clusters.

The Kelley spread tells us how similar, or dissimilar the members of a cluster are to each other. In a chemical context, the Kelley spread for cluster \(m\) can be given by the following formula:

\[
\text{spread}_m = \frac{\left(\sum_{i=1}^{N} \sum_{i=1}^{N} \text{sim}(i,j)\right)}{N(N-1)/2},
\]

where cluster \(m\) contains \(N\) compounds and \(\text{sim}(i,j)\) is the similarity between cluster member compounds \(i\) and \(j\). If all possible values of \(\text{sim}(i,j)\) are between 0 and 1, then \(\text{spread}_m\) will also take a value between 0 and 1. A \(\text{spread}_m\) value of 0 means that the compounds in cluster \(m\) have no features in common, while a \(\text{spread}_m\) value of 1 means that the compounds in cluster \(m\) are identical to each other.

The Kelley distance tells us how dissimilar one cluster is from another. In a chemical context, the Kelly distance between clusters \(m\) and \(n\) can be given by

\[
\text{dist}(m,n) = \frac{\left(\sum_{i=1}^{X} \sum_{j=1}^{Y} \text{sim}(i,k)\right)}{XY},
\]

where cluster \(m\) contains \(X\) compounds, cluster \(n\) contains \(Y\) compounds and \(\text{sim}(i,k)\) is the similarity between compounds \(i\) and \(j\), of clusters \(m\) and \(n\), respectively. If all possible values of \(\text{sim}(i,k)\) are between 0 and 1, then \(\text{dist}(m,n)\) will also take a value between 0 and 1. A \(\text{dist}(m,n)\) value of 0 means that the
compounds in cluster \( m \) have no features in common to the compounds in cluster \( n \), while a \( \text{dist}(m, n) \) value of 1 means that the compounds in cluster \( m \) are identical to the compounds in cluster \( n \). Schuffenhauer et al. clustered the NCI cancer data set and in-house Novartis screening data using molecular frameworks, the scaffold tree and SCINS. As discussed above, this resulted in classes of compounds with similar bioactivity [26]. The Kelley spread was used to evaluate the spread of the different activities of the compounds in a class instead of their structural similarity to each other. The spread of bioactivities in a class decreased as more descriptive scaffolds were used to classify compounds. For example, higher levels of the scaffold tree gave classes with a lower of spread of activity than lower, less-descriptive levels of the scaffold tree. The same was observed for molecular frameworks that gave a low spread of activities compared with graph frameworks that gave a higher spread of activity. These observations can be attributed to the nature of the classes with more descriptive scaffolds; these classes have less variance in chemical structure, and therefore are more likely to have similar activity against the same target. Smaller and less-descriptive scaffold representations contain more structural variance, and are therefore more likely to have different activities. Schuffenhauer et al. also observed that the more descriptive scaffold representations gave a larger number of different classes [26]; this is because a more descriptive scaffold definition will represent a smaller proportion of the data set as illustrated for the ChEMBL data set represented by molecular frameworks and level 1 scaffolds (Table 3.1).

We have discussed several methods for quantifying the distribution of molecules in medicinal chemistry space over scaffolds. Studies utilizing these methods consistently show that medicinal chemistry space is unevenly distributed over scaffolds, with many molecules represented by a small number of common scaffolds. It has been observed for several chemical libraries that this uneven distribution follows a power–law distribution [8,46,47,48]. A power–law distribution implies that large occurrences are rare (such as the small number of well-represented scaffolds in a library) and small occurrences are common (such as many singleton scaffolds in a library). The power–law distribution can be written as

\[
p(x) = kx^{-\alpha},
\]

where \( k \) is the normalization constant of the distribution and \( \alpha \) is the exponent of the power–law and must be greater than 1. The log of this equation

\[
\log p(x) = -\alpha \log x + \log k
\]

gives a straight line with a gradient of \(-\alpha\). Thus, power–law behavior can be identified by plotting the frequency of each occurrence on a log–log scale, for example, plotting the number of scaffolds that represent \( n \) compounds against \( n \) on a log–log scale. In practice, it is often easier to use a rank/frequency plot. A rank/frequency plot is constructed by ranking each occurrence by its frequency, and plotting the rank of the occurrence against its frequency on a log–log scale, for example, ranking scaffolds by their frequency, and plotting the rank of the scaffold against their frequency on a log–log scale. If a distribution follows a power–law
behavior, the rank/frequency plot will give a straight line that takes equation

$$\log p(x) = -(\alpha - 1) \log x + \log \left( \frac{k}{\alpha - 1} \right),$$

and has a gradient of $$-(\alpha - 1)$$. Usually, a data set will not show a linear trend over the entire data set, but over a minimum value of $$x$$. Figure 3.4 shows the rank/frequency plot for the ChEMBL database represented with molecular frameworks and level 1 scaffolds. The central portions of the rank/frequency plots for both the molecular frameworks and level 1 scaffolds form a straight line; therefore the distribution of the ChEMBL database over scaffolds follows power-law behavior for a particular range of scaffold frequencies.

3.4 Visualizing the Scaffold Diversity of Medicinal Chemistry Space

Apart from quantifying the scaffold diversity of medicinal chemistry space, the diversity can be visualized using a variety of techniques. Here we will discuss several methods for visualizing the scaffold diversity of medicinal chemistry space.
Self-organizing maps (SOMs) or Kohonen networks are a type of artificial neural network for reducing complex multidimensional data into a two-dimensional (2D) map and have many applications in chemistry [49–51].

In Section 3.2.2, we discussed the work of Ertl et al. in which almost 600,000 possible ring systems were generated; 780 of these ring systems were present in bioactive molecules [19]. These possible ring systems, represented by various physicochemical properties, were organized into a SOM with a network size of $100 \times 100$ over a period of 200 training iterations. As mentioned previously, the ring systems present in bioactive compounds were present in several distinct activity islands that were sparsely distributed over the map (Figure 3.5).

Tree maps were introduced in 1992 to visualize the directory tree of hard disks [52]. Tree maps visualize hierarchical data structures using a 2D space-filling approach where rectangles or circles represent each leaf of a hierarchical tree. Tree maps have been used to visualize the scaffold diversity of various medicinal chemistry data sets [17]. Figure 3.6 shows an example for version 03 of the ChEMBL database. The database has been represented with level 1 scaffolds; these scaffolds have then been clustered based on their fingerprint similarity. In the tree map, scaffolds are represented by circles; the area of the circle is proportional to the frequency of occurrence of the scaffold. Scaffolds in the same cluster are grouped into a larger gray circle. Circles are also colored by scaffold population, with highly populated scaffolds having a darker color than less-populated scaffolds. This way of visualizing the scaffold diversity of a data set again highlights the uneven distribution of compounds in scaffold space.

Figure 3.5 SOM of 576,556 virtually generated ring systems; 780 of these ring systems have been identified in bioactive molecules and are therefore defined as active scaffolds. The cells of the SOM are color coded by the ratio of active to inactive scaffolds in that cell, showing several activity islands.
Network analysis has been used to investigate and visualize the scaffold diversity of medicinal chemistry space [21,53]. Figure 3.7 Scaffolds are represented with nodes; two nodes are connected with an edge if the scaffolds they represent have a structural similarity over a defined threshold (e.g., ECFP_4 Tanimoto similarity ≥0.70). This produces a network in which similar scaffolds cluster together, providing information on the diversity of scaffolds and their relationships to each other. Subgraphs that are disconnected from the network are examples of related scaffold families. Such subgraphs can be circular or star shaped and represent analogs produced from comparable amounts of modifications from a common parent scaffold, an example of a similarity-based scaffold network. Extended linear subgraphs represent multiple subsequent modifications away from a parent scaffold [21].

Data visualization is an excellent method for representing complex data to illustrate key trends and conclusions. Application of the visualization techniques discussed here clearly reinforces the findings that medicinal chemistry space is represented by a few heavily represented scaffolds and many singleton scaffolds.

Figure 3.6 Tree map of the ChEMBL database version 03 level 1 scaffolds (519,341 compounds, 68,370 level 1 scaffolds). Colored circles represent scaffolds, the area and color of the circles relate to the scaffold frequency, and gray circles represent clusters of scaffolds. Tree maps illustrate the large proportion of singleton scaffolds in the data sets (many small white circles) and the presence of highly populated scaffolds (few large green circles).
3.5 Conclusions

Analysis, quantification, and visualization of scaffold diversity are powerful tools for characterizing and comparing compound libraries. Scaffold representations can
also be used to group molecules into clusters with similar biological profiles. However, such analyses are dependent upon both the definition of the scaffold and the method of diversity assessment. An ideal scaffold definition would be small enough to incorporate multiple substituent vectors but not so large that essentially the whole molecule is defined by the scaffold. Scaffold diversity, in its simplest form, describes the distribution of compound populations over scaffolds; however, more advanced clustering, network methods, and self-organizing maps can be used to analyze the structural diversity of scaffolds in more detail. There is potential for studying scaffold diversity from the perspective of three-dimensional (3D) pharmacophore space in addition to 2D structural analysis.

Analysis of scaffold diversity, defined as the distribution of compound populations over scaffolds, gives a consistent and clear conclusion irrespective of the scaffold definition, visualization, or quantification methods described in the literature that exemplified chemistry space is largely represented by a few highly populated scaffolds. This conclusion is amplified with less-granular scaffolds (e.g., level 1) compared with larger, more defined scaffold definitions (e.g., Murcko frameworks). Natural product libraries may provide an inspiration for diversifying scaffold space but the synthetic chemistry challenge is likely to be high. A challenge for the computational chemistry community is to better describe the vacant medicinal chemistry relevant areas in scaffold space and a corresponding challenge for the diversity-oriented synthesis community is to help populate these holes with synthetically accessible drug-like molecules.

References

10 Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B.,
Scaffold Diversity in Medicinal Chemistry Space

3


References


4
Scaffold Mining of Publicly Available Compound Data
Ye Hu and Jürgen Bajorath

4.1
Introduction

Theoretically, available chemical space has been estimated to consist of more than $10^{60}$ small molecules with at most 30 heavy atoms [1]. However, only a small fraction of this vast chemical space is currently accessible and a very small segment of this accessible chemical space is biologically relevant, that is, contains compounds with specific biological activities [2]. In chemical biology and medicinal chemistry, the exploration of structure–activity relationships (SARs) is of central relevance to characterize bioactive compounds that interact with single or multiple target proteins. SAR exploration aims to reveal how structural features of compounds influence biological activity and, on the basis of this information, guide chemical hit-to-lead and lead optimization efforts.

Traditionally, SARs have been studied on a case-by-case basis in medicinal chemistry, that is, by focusing on one compound series at a time. Modern high-throughput screening techniques yield rapidly increasing amounts of compound activity data [3]. Often, compound sets are sequentially screened against multiple targets or profiled against arrays of related or diverse targets, which produce large amounts of bioactivity data. In addition, compound optimization data become available at increasingly high rates, due to the application of efficient parallel synthesis technologies. During lead optimization, more compounds are typically generated than in earlier years. Taken together, these developments result in the availability of unprecedented amounts of activity data, much of which is ultimately finding its way into the public domain, hence providing a rich source for SAR exploration. For example, increasing amounts of compound data from medicinal chemistry resources are archived in major public repositories, including BindingDB [4] and ChEMBL [5]. Here, compound structures and activity data are extracted from medicinal chemistry literature or patent resources, organized across target families, and curated manually. These databases contain different types of activity measurements, including IC$_{50}$, EC$_{50}$, $K_i$, and/or $K_d$ values for compounds assayed against a wide spectrum of targets. Currently, BindingDB contains 1 010 714 activity records for a total of 421 849 small molecules and 6910 target proteins. In
ChEMBL, 1,487,579 compounds with 11,420,351 activity measurements for 9,844 targets are currently available. Because BindingDB has incorporated a significant amount of ChEMBL data, there is substantial overlap between these databases. In addition to these medicinal chemistry repositories, the PubChem BioAssay database [6] represents the major public resource of compound screening data. It contains more than 700,000 bioassays, including high-throughput screening data sets.

The scaffold concept is one of the most frequently applied concepts in medicinal chemistry and virtual compound screening [7]. Molecular scaffolds (or frameworks) are derived from active compounds by removal of R groups and hence represent their core structures. A number of scaffold analyses have been carried out to associate scaffolds with specific or multiple biological activities [8–12]. In this chapter, we focus on the systematic analysis of SARs in publicly available compound data, with a special focus on molecular scaffolds. A series of studies has been designed to explore different types of relationships between compounds, scaffolds, and biological activities, including the analysis of target or target family selectivity, promiscuity, or scaffold diversity among specifically active compounds. The latter analysis has provided a detailed picture of scaffold-hopping likelihood across different targets, with immediate implications for virtual screening.

4.2 Scaffold Definition

Molecular scaffolds can be derived in different ways. For example, they might be generated by fragmenting compounds on the basis of synthetic or retrosynthetic rules [11,13]. In 1996, Bemis and Murcko introduced a hierarchical molecular organization scheme to analyze molecular frameworks of drugs [14], which has become probably the most popular scaffold concept. Following this definition, atomic property-based Bemis and Murcko scaffolds (BM scaffolds) are obtained by removing all substituents from compounds, while retaining ring systems and linkers between them. Furthermore, BM scaffolds can be transformed into so-called cyclic skeletons (CSKs) by converting all heteroatoms to carbon and all bond orders to one [15]. Hence, a given CSK represents a set of topologically equivalent BM scaffolds. The generation of BM scaffolds and CSKs is illustrated in Figure 4.1. According to the underlying molecular hierarchy, multiple compounds might yield the same BM scaffold and multiple scaffolds the same CSK. This hierarchical structural organization scheme was consistently applied in all studies discussed herein. Furthermore, well-defined activity criteria were consistently applied in the selection of scaffolds. Thus, scaffolds were only extracted from compounds active against human targets for which explicit $K_i$ and/or $IC_{50}$ measurements were available. To provide an up-to-date view of the current data situation, we note that from the most recent release of the ChEMBL database (ChEMBL 15), ~45,500 unique BM scaffolds meeting these criteria are obtained, which originate from more than 124,000 active compounds. These scaffolds correspond to ~18,200
unique CSKs. Thus, a large pool of compound and scaffold data is currently available for systematic SAR exploration and different types of scaffold analysis, as discussed in the following.

4.3 Selectivity of Scaffolds

4.3.1 Privileged Substructures

The idea of “privileged substructures” was originally introduced in 1988 by Evans et al. who observed that many cholecystokinin antagonists contained conserved core structures that were not frequently found in compounds with different activity [16]. Since then the search for target family-privileged chemotypes has continued. Without doubt, the concept of such target family “master keys” is attractive for medicinal chemistry and drug discovery. Privileged substructures are currently understood as structural motifs that are recurrent in ligands of a given target...
family, but not specific for individual targets within the family [9,17]. Accordingly, several privileged substructures have been identified for the G protein-coupled receptor (GPCR) [18] and protein kinase families [19]. However, the existence of truly privileged substructures has often been called into question. For example, Schnur et al. observed that substructures occurring with high frequency in ligands of a particular target family might also act on several other families [10]. Such observations made on the basis of statistical analysis of core structures in ligands of target families can be rationalized if one considers how putatively privileged substructures have typically been introduced. This has mostly been done on the basis of medicinal chemistry knowledge and intuition or frequency-of-occurrence analysis of preselected core structures in ligands of targets of interest [10,19], without considering their global distributions across different target families. Accordingly, a systematic analysis of core structure distributions of active compounds across different target families, without taking subjective criteria into account, was thought to be a meaningful approach to further explore the concept of privileged structural motifs.

4.3.2 Target Community-Selective Scaffolds

In 2010, we reported a comprehensive data mining analysis to explore privileged substructures from a different perspective, that is, by systematically analyzing the activity profiles of compounds at the level of molecular scaffolds [20]. The study was based on all bioactive compounds active against human targets with defined potency measurements available in BindingDB (accessed in September 2009). Compounds were organized into target pair sets. Each set consisted of all available compounds that were active against both targets in a pair. Hence, for targets comprising a family, all possible pairs were generated as a basis for the organization of compounds into pair sets. Only target pair sets that contained at least five compounds were further analyzed. From our collection of 17,929 bioactivity records (compounds might be active against multiple targets), we identified 520 target pairs that involved 6,343 compounds active against 259 human targets. A total of 210 distinct scaffolds that represented at least five compounds were identified. Target pairs were visualized in a network representation in order to analyze ligand-based target relationships (i.e., an edge was drawn between two target nodes if they formed a qualifying pair set). In this network, target communities were formed that mostly represented individual target families. Of the 210 scaffolds extracted from the target pair sets, 206 scaffolds were found to represent compounds exclusively active against a single target community [20]; a rather surprising finding. Although probably influenced by data sparseness, the identification of community-selective scaffolds provided support for the privileged substructure concept. In this context, data sparseness means that as more bioactivity data become available, the number of community-selective scaffolds might be reduced (although this cannot be concluded with certainty). Importantly, studies leading to the identification of community-selective scaffolds considered all
available scaffolds in light of the activity profiles of the compounds they represented and did not rely on frequency-of-occurrence analysis of preselected structural motifs. Hence, no subjective criteria were applied.

Recently, the analysis has been extended to a larger data set, taking into account the growth in compound data since 2010. From ChEMBL release 13 (accessed in June 2012), 33 746 compounds with reported activity against 542 human targets and more than 56 400 explicit Ki measurements were assembled. A total of 363 target pair sets were generated, which consisted of at least five compounds. From these sets, 544 scaffolds were isolated, each of which represented at least five compounds. On the basis of these target pairs, a ligand-based target network representation was designed, as shown in Figure 4.2a. In this network, nodes represented targets that were connected by an edge if they formed a qualifying

![Figure 4.2](target_network.png)

**Figure 4.2** Target network. (a) A compound-based target network is shown. Nodes represent targets and are connected by an edge if two targets share at least one scaffold that is represented by at least five compounds active against both targets. The edge width is scaled according to the number of shared compounds. Target communities formed by at least three targets are numbered from 1 to 21. Targets are colored according to family designations. The remaining targets are colored gray. (b) Eight representative community-selective scaffolds (top) are shown for community 1—8. These scaffolds represent at least five compounds exclusively active against one or more targets within a target community. In addition, five nonselective scaffolds (bottom) are shown, which are associated with compounds active against two or five target communities. For each scaffold, the target community designation is reported.
Nodes were colored according to their target family designations and the edge width was scaled according to the number of shared compounds. From this network, 21 distinct communities emerged, which consisted of at least three targets. The color distribution of the nodes revealed that the majority of the communities comprised targets belonging to individual target families. For example, community 2, 3, and 4 exclusively contained GPCR, carbonic anhydrase, and metalloprotease targets, respectively. Community 1, 7, and 12 were exceptions because they were formed by targets originating from two or four different families. Of the 458 scaffolds that were present in these 21 communities, 453 scaffolds were found to be community selective. Representative community-selective scaffolds and the remaining five nonselective scaffolds are shown in Figure 4.2b. As can be seen, chemically similar scaffolds displayed different selectivity across target families. For example, the phenylcyclohexane scaffold selective for community 1 (i.e., the GPCR family) and the biphenyl scaffold selective for community 3 (i.e., carbonic anhydrases) only differed by the aromatic character of one six-membered ring. Moreover, scaffolds selective for community 6 (i.e., the serine protease family) and 7 (mostly GPCRs) were chemically distinct but represented the same topology.

Thus, over the course of two years, the number of community-selective scaffolds we identified more than doubled. This implies that compounds containing new scaffolds are evaluated at high rates. However, it is unlikely that these compounds are only tested on individual target families. Although we would not expect that all of the scaffolds we identified are truly target family selective, given the implications of data sparseness, the increasing number of observed community-selective scaffolds further supports the notion of privileged substructures.
4.3.3 Target-Selective Scaffolds

We also analyzed the selectivity profiles of compounds represented by the 206 originally identified community-selective scaffolds in order to determine whether target-selective scaffolds might also exist. For each compound in a target pair set, its target selectivity (TS) value was calculated as the logarithmic potency difference for the target pair. For each community-selective scaffold active against a given target, all compounds containing this scaffold were pooled, TS values were calculated for this target, and the median TS value was determined. The median TS served as a measure for the selectivity of a scaffold for a given target. The majority of the community-selective scaffolds displayed a tendency to yield selective compounds for one target over others, that is, compounds a scaffold represented were preferentially – but not exclusively – highly potent against one target in a community [20]. However, in most instances, scaffolds with strong target-selective tendency only corresponded to one or two compounds. Therefore, truly target-selective scaffolds could not be confirmed at this stage [21]. Thus, the assessment of target selectivity was complicated by data sparseness within communities. Before postulating target specificity of a given scaffold, one would like to see activity of corresponding compounds against larger numbers of targets. It should also be noted that scaffolds with strong target-selective tendency must not necessarily be community selective. They could be highly potent against a given target over others belonging to different families.

4.4 Target Promiscuity of Scaffolds

4.4.1 Promiscuous BM Scaffolds and CSKs

Target selectivity and promiscuity mark opposite ends of the compound activity spectrum. Over the past decades, the traditional target specificity paradigm has dominated pharmaceutical research. This means that the discovery and optimization of compounds with apparent target specificity has been the major goal of drug discovery efforts. However, in recent years, the specificity paradigm has been under reconsideration because it has been shown that many pharmaceutically relevant compounds and drugs elicit their therapeutic effects by acting on multiple targets [22–25]. Such promiscuous compound behavior provides the conceptual basis for the notion of polypharmacology that is beginning to influence drug discovery efforts [26,27]. For the treatment of many chronic or infectious diseases, target-specific compounds continue to be highly preferred, but for other applications such as the treatment of many forms of cancer, multitarget activities of drugs are highly desirable (as exemplified, e.g., by promiscuous protein kinase inhibitors that are in clinical use). Given the increasing relevance of multitarget activities, it has been of
high interest to evaluate compound promiscuity on a large scale. We have extended this analysis to molecular scaffolds and carried out a systematic survey based on all BindingDB and ChEMBL compounds that were active against human targets with at least 1 μM potency (accessed in May 2010) [28]. Compounds comprising a total of 458 target sets belonging to 19 target families were collected. These target sets consisted of 34,906 compounds that yielded 13,462 unique BM scaffolds. From compounds represented by each scaffold, target annotations were assembled and an activity profile was generated. On the basis of these scaffold-centric activity profiles, a total of 435 scaffolds were found to have activity against targets from at least two different families. Of these, 83 scaffolds were active against 3–13 target families and were categorized as “promiscuous scaffolds.” These promiscuous scaffolds represented 33 distinct CSKs that were considered “promiscuous chemotypes.” Representative promiscuous chemotypes active against targets from 4 to 15 families are shown in Figure 4.3a. These scaffolds were structurally diverse, as they contained one to eight ring structures, hence reflecting a remarkable diversity.

![Figure 4.3](image)

**Figure 4.3** Promiscuous chemotypes. (a) Fourteen representative promiscuous CSKs that are active against at least four target families are shown. For each CSK, the number of target families and the number of corresponding promiscuous BM scaffolds are reported. For instance, “15|5” indicates that the CSK is active against targets belonging to 15 target families and represents five promiscuous BM scaffolds. (b) Scaffold—target family relationships are shown in a bipartite network for dicyclohexylmethane. Gray circular nodes represent promiscuous scaffolds covered by the CSK and white rectangular nodes represent target families. An edge is drawn between a scaffold and a target family if compounds represented by the scaffold are active against targets belonging to the family. (c) A target-based scaffold network is shown. Nodes represent six promiscuous BM scaffolds covered by dicyclohexylmethane. Activity profiles of these scaffolds are compared. Two nodes are connected by an edge if they share one or more targets. For each scaffold, the number of targets and families is reported. For example, “4|10” means that the scaffold is active against 10 targets belonging to four target families.
variation in structural complexity. Data incompleteness is also likely to affect the assignment of promiscuous scaffolds. However, different from the situation with community-selective scaffolds, more scaffolds are expected to reveal promiscuity as more activity annotations become available. On the basis of currently available data,
there are already more than 80 promiscuous scaffolds and more than 30 promiscuous chemotypes available, which can be used to explore the design of promiscuous compounds.

4.4.2 Scaffold–Target Family Profiles

For each of the 33 promiscuous CSKs, relationships between the scaffolds it covered and their target families were determined. Figure 4.3b shows a bipartite scaffold–target family network for the CSK dicyclohexylmethane that displayed 31 scaffold-family relationships, representing the largest number of relationships observed among 33 CSKs. In this network, circular nodes were BM scaffolds corresponding to a given CSK and rectangular nodes were target families. An edge was drawn between a scaffold and a target family if compounds represented by the scaffold were active against target(s) of the family. The network revealed different relationships between target families and topologically equivalent scaffolds. In this example, the scaffolds were active against three to eight target families, indicating different degrees of promiscuity. In addition, scaffold (with ID) 11 050 displayed an overlapping family profile with scaffold 8251, whereas the profile was distinct from 15 126. Activity profile comparisons of these six topologically equivalent scaffolds were further carried out on the basis of their target annotations, as shown in Figure 4.3c. In this target-based scaffold network, nodes represented scaffolds that were connected if they shared one or more target annotations. Compounds representing these six scaffolds were active against 8–29 different targets and shared 1–6 of these targets with at least 2 other scaffolds. Thus, the analysis revealed different activity profiles of closely related scaffolds. For example, four scaffolds (11 050, 12 191, 1930, and 8251) only differed by a single atom in the linker. However, their activity profiles were rather diverse, that is, compounds representing these scaffolds were active against 10–29 targets. Taken together, these findings revealed that closely related topologically equivalent scaffolds represented compounds with often very different activity profiles.

4.4.3 Promiscuous Scaffolds in Drugs

The global distribution of promiscuous BM scaffolds in bioactive compounds and drugs was also determined. The set of 83 promiscuous scaffolds discussed in Section 4.4.1 represented ~6% of all active compounds. These 83 scaffolds were mapped to 1247 approved drugs taken from DrugBank [29] (accessed in May 2010). A subset of 39 scaffolds corresponding to 17 unique CSKs was detected in 215 (~17%) of these drugs, thus indicating a significantly higher proportion of promiscuous scaffolds contained in drugs than in bioactive compounds [28]. Because drugs originate from the pool of bioactive compounds, target-promiscuous compounds might be enriched during the drug development process, which must essentially be attributed to reasons of efficacy observed during preclinical testing.
Otherwise, it would be difficult to explain why drugs containing promiscuous scaffolds were put forward, although they departed from the target specificity paradigm that governed drug discovery efforts during the time of their development (as discussed earlier).

4.5 Activity Cliff-Forming Scaffolds

4.5.1 Activity Cliff Concept

A different type of molecular scaffold analysis has been focused on activity cliffs that are generally defined as pairs or groups of structurally similar compounds having large potency differences and which represent the most discontinuous region in activity landscapes [30,31]. Activity cliffs are of high interest in medicinal chemistry because they often reveal SAR determinants [31], given their “small structural change–large biological effect” phenotype, and are typically studied on the basis of pairwise compound comparisons. We have asked the question whether scaffolds might exist, which have an intrinsic propensity to form activity cliffs across different targets. Therefore, we have carried out a systematic analysis of compounds with activity against multiple targets sharing the same scaffolds [32].

4.5.2 Multitarget Cliff-Forming Scaffolds

From ChEMBL and BindingDB (accessed in January 2010), a total of 47,004 unique compounds that were active against 880 human targets and yielded 17,784 unique BM scaffolds were assembled for this analysis. To quantitatively evaluate the propensity of a scaffold to form activity cliffs, a potency-based scaffold discontinuity score (PScS) was designed as follows:

\[ PScS(s) = \frac{\sum(|p_i - p_j| \times sim(i,j))}{|ij|}. \]

Here, \(|p_i - p_j|\) is the absolute potency difference of a pair of compounds i and j represented by the scaffold s, sim(i,j) is the Tanimoto similarity [33] of compounds i and j using 166 MACCS structural keys1 as descriptors, and \(|ij|\) is the number of all possible compound pairs sharing the scaffold s. Accordingly, the scoring scheme involves systematic pairwise structural similarity and potency comparisons of compounds containing the same scaffold. Initially, calculated (raw) scores were normalized with respect to the distribution of scores for all scaffolds to obtain final scores between 0 and 1. A score close to 1 indicated the presence of significant activity cliffs within a set of compounds [32]. For each scaffold, all corresponding

compounds were first collected for all targets they were active against and a global PScS score was calculated. Scaffolds were prioritized, which produced a high global discontinuity score (i.e., greater than 0.8) and which were represented by at least three compounds active against at least two targets. These selection criteria resulted in a set of 212 BM scaffolds. For each of these scaffolds, the score was recalculated on a per-target basis in order to further prioritize scaffolds that had a high propensity to form activity cliffs for different targets. Hence, scaffolds were prioritized, which produced a score of greater than 0.8 for at least two targets. On the basis of this criterion, 103 scaffolds were identified, which represented compounds forming multitarget activity cliffs. Exemplary scaffolds are shown in Figure 4.4a. These scaffolds revealed different degrees of structural complexity and diversity.

The activity profiles of individual multitarget cliff-forming scaffolds were further analyzed in scaffold-based target networks, shown in Figure 4.4b. In these networks, nodes represented targets that a given cliff-forming scaffold was active connected by an edge if they share compounds containing the cliff-forming scaffold. The nodes are colored according to the discontinuity scores (red, if the score is higher than 0.8). Four exemplary compounds are shown and their potency values (pKi) for individual targets are reported. Target abbreviations: Akt1, RAC-alpha serine/threonine-protein kinase; Akt2, RAC-beta serine/threonine-protein kinase; Akt3, RAC-gamma serine/threonine-protein kinase; MSK1, mitogen- and stress-activated protein kinase 1; ROCK1, Rho-associated protein kinase 1; RSK1, ribosomal S6 kinase 1.
against. Two nodes were connected if they shared compounds containing this scaffold. In addition, nodes were colored according to target-based discontinuity scores. In this example, the network was generated for a scaffold with a global discontinuity score of 0.84. The scaffold represented a total of 10 compounds that were active against six targets and displayed a high propensity to form activity cliffs for four targets. As shown in Figure 4.4b, structural modifications leading to potency changes could be rationalized.

4.6 Scaffolds with Defined Activity Progression

4.6.1 Activity Profile Sequences

Recently, we have introduced the concept of activity profile sequences that accounts for the progression of compound activity in target space. The approach was applied to extract SAR information from analog series, with multiple target annotations [34]. From ChEMBL (accessed in January 2011), we collected 66,796 compounds active against 825 human targets with at least 10 μM potency. For each compound, its activity profile was generated by assembling all available target annotations. Compounds were then organized on the basis of activity profiles. Accordingly, a total of 959 unique activity profiles containing at least 5 compounds were extracted from ~67,000 compounds. Each profile covered a set of unique compounds. In the
next step, activity profiles were compared in a pairwise manner to determine subset relationships, as illustrated in Figure 4.5a. Finally, profile pairs were organized into profile sequences in the order of increasing numbers of targets. Therefore, any two adjacent profiles along the sequence differed by the addition of exactly one target.

(a) activity profiles

<table>
<thead>
<tr>
<th>Profile Pair</th>
<th>Profile Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_A$</td>
<td>$T_A T_B$</td>
</tr>
<tr>
<td>$T_A T_B$</td>
<td>$T_A T_B T_C$</td>
</tr>
<tr>
<td>$T_A T_C$</td>
<td>$T_A T_B T_C$</td>
</tr>
<tr>
<td>$T_A T_B T_C$</td>
<td></td>
</tr>
</tbody>
</table>

(b) Representative conserved scaffolds are extracted from profile sequences of length 4, each of which represents more than 20 active compounds. For each scaffold, the number of sequences of length 4 that it contains and the number of compounds the scaffold represents are reported. For instance, “6/45” indicates that a scaffold occurs in a total of six different profile sequences of length 4 and represents 45 compounds. (c) A representative profile sequence of length 4 involving four adenosine receptors is shown. Exemplary compounds containing the conserved scaffold (red) are shown along the profile sequence.

Figure 4.5 Activity profile sequences and conserved scaffolds. (a) How to generate activity profile sequences is illustrated. On the basis of a set of unique activity profiles, subset relationships are determined for all profile pairs that differ by only one target. The profile pairs are then organized into a sequence in the order of increasing numbers of targets. (b) Sixteen representative conserved scaffolds are extracted from profile sequences of length 4, each of which represents more than 20 active compounds.
4.6 Scaffolds with Defined Activity Progression

(Figure 4.5a). A total of 671 profile sequences of length 2–6 were identified. The majority of these sequences had a length of 2–4 and involved targets from single or multiple families. In addition, 36 sequences consisted of five or six profiles exclusively involving targets from the same family [34].

4.6.2 Conserved Scaffolds

We further analyzed the scaffold composition of activity profiles along the profile sequences. Although each profile was assigned a set of unique compounds, a total of 405, 51, and 26 scaffolds were found to be conserved across all activity profiles in sequences of length 2, 3, and 4, respectively. These conserved scaffolds represented analog series with activity progression along the sequences across all profiles and were diverse in complexity and topology. Figure 4.5b shows 16 conserved scaffolds that were identified from profile sequences of length 4 and represented more than 20 active compounds. In addition, an exemplary profile sequence of length 4 is given in Figure 4.5c. This sequence involved four adenosine receptors. Structural modifications of compounds represented by the same scaffold led to progression of activity from single toward multiple targets.
Another intensively studied topic in medicinal chemistry and virtual screening is scaffold hopping that refers to the identification of compounds with distinct core structures having similar activity. The search for scaffold hops represents the primary goal of ligand-based virtual screening. In general, it is unclear how difficult scaffold hopping exercises might be. Depending on the target, it might be more or less challenging to identify structurally distinct compounds that are active. One way to estimate the difficulty of successfully identifying such structures is to analyze the scaffold diversity among compounds with target-specific activities. The more structurally diverse scaffolds are already found in specifically active compounds, the easier it should be to identify additional ones. The already existing scaffold diversity can thus be used as a global measure for the degree of difficulty involved in identifying structurally distinct active compounds. In other words, structural diversity among known ligands provides information about the ability of a target to bind small molecules. For targets that are “small molecule friendly,” new structures are easier to identify than for difficult small molecule targets, corresponding to greater scaffold hopping potential. In order to analyze intrinsic scaffold-hopping ability on the basis of available ligand information, we have carried out a large-scale analysis of scaffold diversity on a per-target basis [35]. From ChEMBL and BindingDB (accessed in May 2010), 502 target sets that comprised \( \approx 26,700 \) compounds active against 19 different families were assembled. Each target set contained at least five compounds with at least 1 \( \mu \text{M} \) potency and at least two different BM scaffolds. We first determined the number of distinct scaffolds that were present in each target set. A total of 354, 42, and 28 target sets consisted of between 5 and 49, between 50 and 99, and at least 100 topologically distinct scaffolds, respectively. Therefore, the range of 5–49 scaffolds represented the average scaffold diversity across currently available targets [35]. Figure 4.6 shows the scaffold distribution within individual target families. As can be seen, several well-known pharmaceutical target families were characterized by high scaffold diversity, including several protease families, kinase families, and GPCRs. The targets belonging to these families were chemically well explored, revealing a high degree of permissiveness to structurally diverse small molecules, corresponding to significant scaffold hopping potential.

### 4.7.2 Structural Relationships between Scaffolds

In addition to the general assessment of the frequency of scaffold hops across current targets, we also systematically investigated structural relationships between
scaffolds [36]. On the basis of structural features of BM scaffolds and CSKs, two types of structural relationships were determined:

i) a BM scaffold is a substructure of another one.
ii) two different BM scaffolds share the same CSK.

These structural relationships were explored for scaffolds in more than 500 target sets. Although the most generic scaffold (i.e., the benzene ring) was excluded from the comparison, structural relationships were detected in ~87% of all target sets. Surprisingly, for more than 20% of the target sets, all scaffolds were found to be involved in one or both types of structural relationships. In total, more than 80% of the scaffolds were structurally related [36]. In addition, we have also carried out an analysis of structural relationships between scaffolds from a set of 1586 clinical trial compounds, 2980 registered or launched drugs, and 50 000 compounds that were randomly collected from ZINC [37]. These scaffolds were not organized into target sets. More than 90% of these scaffolds were involved in structural relationships [36]. This high rate was rather unexpected. Although a large number of scaffolds were found for many pharmaceutically relevant targets, the majority of these scaffolds were structurally related. This might suggest that biologically relevant scaffold space is smaller than one would expect. Many targets displayed high intrinsic scaffold hopping potential. However, the search for structurally unrelated scaffold hops should be rather challenging, as indicated by our findings.

Figure 4.6 Scaffold distribution. The distribution of scaffolds for 502 target sets belonging to 19 families is shown in box plots. Each box plot reports the smallest value (bottom line), the first quartile (lower boundary of the box), median (thick line), the third quartile (upper boundary), and the largest value (top line). Target family designations are also given.
4.7.3 Scaffold Hopping in Virtual Screening

Because scaffold hopping is the primary goal in ligand-based virtual screening, we have also conducted a systematic analysis of scaffold hopping characteristics using two-dimensional (2D) molecular fingerprints [38] that are among the most popular similarity search tools. Different from the other investigations discussed herein, this study was not a data mining effort, but rather a benchmark investigation. First, 17 different compound activity classes were designed, which consisted of compounds containing 11–46 different BM scaffolds. In each activity class, every scaffold was represented by exactly five compounds that had a potency value ($K_i$ or $IC_{50}$) of at least 1 µM and consisted of a maximum of 50 nonhydrogen atoms. In addition, for each compound, the scaffold had to be at least of the same size as its combined R group set. In total, 1675 compounds corresponding to 334 unique BM scaffolds were assembled [38]. For similarity searching, reference compounds and test sets were organized such that each correctly detected active compounds represented a scaffold hop. Five representative 2D fingerprints of different designs were selected for this study, including MACCS (166 structural fragments), ECFP4 and ECFC4 (representative topological fingerprints) [39], and TGT and GpiDAPH3 (encode 2D pharmacophore features). For each fingerprint, the distribution of Tanimoto coefficient (Tc) values [33] was determined to investigate the Tc range where scaffold hops might occur. One of the key findings of this benchmark study was that each of these five fingerprints showed at least limited scaffold-hopping ability. In addition, we also showed that it was essentially impossible to define generally applicable Tc threshold values for scaffold hopping [38]. However, several practical guidelines were formulated for the use of 2D fingerprints in the search for new scaffolds. For example, for MACCS, a Tc threshold of 0.8 typically yielded database selection sets of around 100 compounds, containing only a small number of scaffold hops, which were notably enriched within the MACCS Tc range of 0.4–0.6. However, more than 50% of the background database compounds were detected within this Tc range. Thus, for similarity searching, this threshold would not be meaningful. For ECFP4, which displayed the overall highest scaffold hopping potential in this study, a threshold of 0.4 was found to contain scaffold hops in database selection sets of about 100 compounds. This was the case because the database Tc value distribution of ECFP4 was shifted to much lower values compared to, for example, MACCS. Taken together, the calculations revealed that small numbers of scaffolds could consistently be detected in relatively small database selection sets. However, significant scaffold recall could not be facilitated using 2D fingerprints because they typically also recognized large numbers of database compounds at Tc threshold values that identified at least a third of different active scaffolds available in the database.

4.8 Conclusions

In this chapter, we have introduced a variety of data mining approaches to systematically explore different types of SAR characteristics associated with scaffolds and to study different types of relationships between scaffolds. Scaffold-centric data mining efforts have focused on publicly available compound activity data. A hierarchical organization scheme has been consistently applied to generate scaffolds from bioactive compounds.

We have revisited the concept of privileged substructures and investigated compound selectivity at the level of target families and individual targets. Although current compound data is still sparsely distributed, more than 400 BM scaffolds were identified, which were selective for individual target communities. Many of these scaffolds displayed a tendency to be target selective, but there were still too few active compounds available to propose truly target-selective scaffolds. By contrast, a total of 83 scaffolds corresponding to 33 CSKs were found to be promiscuous in nature, that is, compounds the scaffolds represented were active against targets belonging to three or more target families. Promiscuous scaffolds sharing the same topology were often chemically closely related but displayed different activity profiles. It has also been observed that these promiscuous scaffolds were enriched in approved drugs, which directly relates to increasingly observed polypharmacological drug behavior.

Furthermore, approximately 100 scaffolds were detected to form activity cliffs across multiple targets with high propensity. These scaffolds represented analog series with chemical modifications that caused significant potency changes for different targets. In addition, other scaffolds were identified, which had conserved activity profiles along profile sequences that reflected activity progression in target space. Structural modifications of these scaffolds led to different degrees of target promiscuity.

Finally, we have generally assessed the scaffold hopping potential for current drug targets. Although the majority of these targets had ligands containing a large number of scaffolds, most of these scaffolds were structurally related, that is, they were involved in substructure relationships and/or shared the same topology. These findings demonstrated that active scaffolds were abundant across many targets, but that structural diversity among scaffolds was lower than anticipated, which has implications for virtual screening. Hence, while it should be relatively easy to facilitate scaffold hopping in many instances, the identification of chemically unrelated scaffolds might be difficult.

In summary, publicly available compound data that mainly originated from the medicinal chemistry literature have provided a valuable resource for large-scale scaffold and SAR analysis. On the basis of the scaffold concept, different SAR features have been explored by taking statistical criteria, target activity/selectivity, promiscuity, potency distributions, and structural relationships into account. Different sets of scaffolds representing various SAR characteristics are thought to provide useful information for compound design and optimization.
References


5
Exploring Virtual Scaffold Spaces
William R. Pitt and Boris Kroeplien

5.1
Introduction

5.1.1
Virtual Chemistry

The exploration of virtual chemistry is bounded only by the human imagination and the capabilities of computers. One can dream of any sort of structure, containing any of the elements, bonded together in unlimited ways. However, this is the realm of artists rather than applied scientists. At least in the field of drug discovery, most like to restrict ideas to those that comply with the known laws of physical organic chemistry. Yet this restraint still leaves an effectively unlimited number of possibilities. However, not all chemical structures are likely to have therapeutic potential. First of all they must be either synthesized or extracted from Nature. It has often been said among chemists that any compound can be made, given enough time. However, medicinal chemistry is tied to the feedback loop of design, synthesis, and testing – a cycle that must be repeated thousands, even hundreds of thousands of times before a drug is produced. Only with good reason will a medicinal chemist spend more than a month trying to make a single compound. Random virtual compounds that are overly complicated will be of very little interest. Complex molecules may also have a very low chance of binding to a selected target protein [1]. So, there are constraints on scope of all virtual libraries and especially those created for use in drug design.

5.1.2
Chemical Space

One can think of a collection of chemical structures floating in abstract space, with those that are similar in some way or that share certain characteristics, positioned close together. There have been efforts to visualize this kind of space in a publishable form, and some of these will be mentioned later. However, the exploration of chemical space, in context of this chapter at least, is simply the
process of constructing a particular subset of structures in a computer, selected from an infinite number of possibilities, for a particular purpose. If this purpose is to provide ideas for direct consumption by a synthetic chemist, some quality control is needed. The chosen subset should be attractive from a number of perspectives. These include chemical stability, synthetic tractability, drug-likeness, novelty, suitability for relevant biological targets, and natural product-likeness. The prediction of each of these properties is the subject of extensive research and largely beyond the scope of this chapter. When producing large virtual libraries, simple physical properties filters are usually applied because they are quick to calculate, well understood, and can dramatically reduce the chemical space under consideration. Often these calculated properties are additive and can be applied prior to enumerations from constituent fragments, for example, molecular weight and SlogP [2]. Another type of filter that can be applied is the removal of reactive and toxic substructures [3]. A conservative approach to synthetic tractability is to employ known reactions schemes and available reagents when constructing virtual libraries. This subject will be discussed in Section 5.4. All these considerations provide boundaries to the chemical space exploration. A way to reduce the space further, down to a manageable size, is to stick to very small structures and thereby exploit the fact that the number of potential compounds grows exponentially with the number of constituent atoms [4].

5.1.3 Scaffold Definition

As well as finished products, one can also explore the potential options of each of the building blocks of which they are composed. In the context of combinatorial or parallel synthesis library design, one can consider these building blocks to be divided into scaffolds and attached R groups [5]. However, there is no commonly agreed definition of what a scaffold is in the context of medicinal chemistry. In the Scaffold Hunter [6] documentation, scaffolds are defined as the “underlying molecular frameworks that serve as simplified representatives for classes of similar molecules.” This definition encompasses the maximum common substructure (MCS) of a series of molecules, and also smaller substructures within it. However, this definition can only really be applied to a preexisting set of molecules, where pharmacophoric constraints and synthetic routes have created patterns in the structure–activity relationship (SAR). For virtual libraries no such SAR exists. Therefore, a different definition must be adopted. Ring systems, with or without connecting linkers within a structure, are often considered to be scaffolds, cf. Bemis and Murcko frameworks [7]. Where no rings are present, the definition of the scaffold becomes even less well defined. In the case of peptides, the main chain is the obvious choice. Similarly, in macrocycles, side chains often project off a cyclic scaffold. In this chapter we will keep an open mind and include any virtual chemical space explorations that could be relevant to the replacement of conserved substructures within an existing chemical series or the design of a library, based upon a common core. We will divide exploration into comprehensive enumeration and iterative stochastic approaches (Figure 5.1).
The Comprehensive Enumeration of Parts of Chemical Space

One approach to generating a virtual chemical library is to exhaustively enumerate all possible alternatives within well-defined bounds. Of course one must choose these bounds with care in order to keep the number of structures manageable. When this is done, it allows the comparison of what could be made with those that have actually been made and proven useful for drug discovery. Databases of compounds and their activities such as ChEMBL [8], DrugBank [9] and PubChem [10], which are freely available to the public, greatly facilitate comparisons with bioactive molecules. Databases of commercially available compounds such as ZINC [11] and ACD [12] provide another source of known chemistry. There are several studies of this kind and in each case the virtual libraries were constructed without reference to known chemical reactions.

5.2.1 Fragments

The most ambitious work of this kind has been done by Reymond and coworkers [4,13–15]. In their most recent paper, they describe the enumeration of 166 billion organic small molecules with up to 17 atoms. The resulting structures have molecular weights lying mostly in the 220–260 Da range. All possible graph topologies were created, and then individual graphs were converted into molecular structures by assigning elements to nodes and bonds to edges in systematic combinations.

Figure 5.1 The different ways to explore chemical space by (a) iterative stochastic search or (b) comprehensive enumeration. Each circle represents a chemical structure. Blue circles are those that are selected as fit for purpose, or for further modification.
This huge task took 100,000 CPU hours. The resulting library is called GDB-17. Many filters were applied to prevent the creation of structures thought likely to be unstable or containing unwanted functionality. The final library was then compared to bioactive molecules from PubChem, ChEMBL, and DrugBank with up to 17 atoms. They found that the proportion of structures containing nonaromatic heterocyclic rings was about five times larger in GDB-17 than in the known molecules, but the fraction containing aromatic rings was much lower. A larger proportion of structures with three or more hydrogen bond donors, or a negative SlogP were also discovered in GDB-17. It was also relatively rich with structures with more than one quaternary center and four or more stereo centers. An unexpected result was that there are many more tricycles in the GDB-17 than in bioactive molecules. This analysis points to bioactive molecules being simpler and flatter in structure in general than those created systematically using an algorithm. Since natural products are known to be often very complex in structure, but relatively few in number in the databases of bioactive molecules, it is probably safe to assume that these trends apply mainly to man-made molecules. Among the reasons for this preference for simple flat molecules are probably the advances in sp²-coupling reactions that have made the synthesis of these molecules relatively easy [16].

5.2.2

Ring Systems

Two other studies applied a similar approach to heteroaromatic ring systems. Ertl et al. [17] enumerated a virtual library of half a million mono-, bi-, and tricyclic ring systems. They found a low diversity of simple aromatic scaffolds in bioactive molecules. This prompted them to investigate whether the low diversity in known aromatic scaffolds was due to (i) the need to be compatible with biological macromolecules, (ii) the limitations of synthetic accessibility, or (iii) the limited time we have spent exploring the vastness of chemical space. After plotting all the ring systems on a Kohonen map, they found that bioactive molecules clustered together into roughly six “activity islands.” While stating that this clustering was due to a complex interplay of multiple factors, nevertheless they concluded that molecular weight and the number of heteroatoms were important factors. They also concluded that (iii) was undoubtedly true.

The second study that focused on heteroaromatic ring systems was carried out by Pitt et al. [18]. In this work a virtual library of about 25,000 mono- and bicyclic aromatic ring systems called VEHICLe was constructed from combining functional group building blocks. The database was separated into two sets by whether they were found in databases of known molecules or not. The differences between the two were then divined with a machine learning procedure using small substructures as inputs. When all of the ring systems were fed back into the “machine” in prediction mode, some unknown examples were predicted to belong to the known set. When further literature searches were done, it was found that many of these
had in fact been made before. The remainder was predicted to be more likely to be synthetically tractable. The authors found that known heteroaromatic ring systems, and especially those appearing in drugs, were heavily biased toward those with a low number of heteroatoms.

In all three studies, relatively complex structures were found to be underrepresented in man-made bioactive molecules. In the first shape complexity was found to be limited, and in all three the number of heteroatoms was found to be reduced. These trends are now well recognized and lab-based academics are working to create real libraries of more complicated and diverse structures, often inspired by natural products [19]. However, given these results, limiting the complexity of virtual compounds and scaffolds is a valid strategy to reduce the number of options to a manageable number and to increase their chances of being synthesized and of having biological activity against a given biological target.

A study by Tu et al. [20] extracted preenumerated aromatic ring systems from GDB [14,15] database and elaborated them further. This set of structures was then put together with a variety of commercial software packages into a system called NEAT, designed to help users to find similar aromatic ring systems to a query. Shape, electrostatic potential (ESP) calculated using QM-derived point charges, hydrogen bonding strength, dipole moment, and a number of other descriptors could be used to search and compare ring systems. This facilitated the search through chemical space for replacement aromatic scaffolds.

5.2.3 Reagents

Exhaustive enumeration can also be used to identify areas of chemical space that have not yet been exploited. A good example for this is the application of virtual libraries to reagent sets for use in medicinal chemistry by workers at AstraZeneca (AZ). The first paper on the subject by Kettle et al. [21] did not in fact rely upon enumeration. Instead they started from a database of two million compounds assembled from the literature and patents. Then they used SMIRKS [22] to fragment the molecules and retrieve any theoretical cyclic amines they contained. From this they built a virtual library of a 16 000 unique cyclic secondary amines that had all appeared in at least five of the literature compounds. They then filtered out all reagents that were available commercially or in-house at AZ. Before deciding which amines to get custom-made and add to the AZ collection, they also applied filters for molecular properties and undesirable functionality. In the end they identified 930 unique interesting cyclic secondary amines. This work can be seen as on the boundary between virtual and real compounds. Most of these amines will have existed before this study, and will have actually been used by the original authors to synthesize the molecules in the first place. But for some of the amines created in silico, the compounds may well have been synthesized by different routes. And so these amines resulting from the SMIRKS manipulation are true virtual compounds.
In a second publication [23], the same authors took this to the next level and worked in truly virtual chemical space. They used their algorithm BOOMSLANG [24] to exhaustively enumerate the virtual library of all six-and five-membered aromatic rings attached to 11 commonly used functional groups (50612 molecules). They then applied substructure and property filters to remove compounds with high lipophilicity, concerns about ring stability, or secondary functional groups, which might interfere with using the molecules as reagents. This reduced the size of their library to 5759 potential reagents.

After cross-referencing these reagents against the ones available in-house at AZ or commercially in the ACD [12], some reagents were purchased to augment AZ’s collection. The novel reagents, which existed neither at AZ nor commercially, were prioritized by medicinal chemists, and a selection of 97 reagents was synthesized at a contract research organization. Even though these 97 new reagents are a very small percentage of the original virtual library, they have added a significant breadth to AZ’s reagent set in this area.

5.3
The Iterative Generation of Virtual Compounds

If one wishes to explore possible structures for compounds with a molecular weight above 300 Da or so, complete enumeration becomes impractical. Thankfully, many sophisticated algorithms exist for efficiently searching large, multi-dimensional spaces. Virtual libraries can be constructed by continually modifying one or more starting structures and then selecting the best progeny to be taken forward for further modification or as finished products.

5.3.1
Transformations

During the lead optimization process, chemists continually alter compounds in various ways in order to improve results in assays of activity, off-target activities, physical properties, metabolic stability, and so on. The subject of this book is the replacement of scaffolds. This can be achieved in a number of ways apart from the most obvious direct bioisosteric replacement. The core can also be gradually mutated until it is sufficiently different. This can be done by the addition, replacement, or (less commonly) the removal of a functional group. There are also cyclization and the replacement of rings with intramolecular hydrogen bonding groups.

5.3.2
Manual Selection of Chemical Modifications

There have been numerous attempts over the last decade to capture common transformations based upon known experience and historical SAR data; one example is the work by Raymond et al. [25]. Essentially, chemists have always been
exploring the chemical space of all the compounds that they know or can find out how to make. In the last few years, these studies have been augmented by matched-pair analyses, which highlight potentially useful changes in general [26–29]. Once these tricks of medicinal chemists are encoded in a language computer programs can readily use, such as SMIRKS [22], then expert systems can be developed. One such system is called Drug Guru [30]. This system includes rules for functional group transformations and molecular framework modifications. They contain changes that are designed to improve solubility and other desirable properties. These transformations can be applied, under user control, to an input molecule. The selection of interesting progeny is left largely to the user in the expectation that the results will widen the number of modifications under consideration for synthesis. A similar approach is called the Molecule Evolutator [31]. This system is designed to generate new molecules over several rounds of directed evolution, with user selection as the fitness function.

5.3.3 Analog Generators

A conceptually simple extension of the iterative transformation of molecular structure under user control is to use similarity to the original molecule as a fitness function. If shape and electrostatic rather than topological similarity are used, this can be applied as an automated scaffold hopping procedure, albeit with the significant added complication of 3D conformer generation but also greater opportunity for novelty. This is the approach taken by the software developed by Ertl and Lewis, called IADE [32]. They used a fragment bioisostere replacement strategy with the Cresset field similarity [33] as the fitness function. The next level in sophistication is to take multiple active molecules and try to find novel structures that are equally similar to all of them. This concept was explored by Brown et al. [34] for two target molecules, using a genetic algorithm to generate “median” molecules.

5.3.4 Inverse QSAR

Inverse QSAR is a technique that was first developed in the late 1980s [35] but has enjoyed a resurgence in the literature recently. It starts with the construction of a QSAR model built from a set of known actives. The process is then inverted and novel structures of compounds are constructed, which are predicted to be active by the model. These virtual compounds can be drawn from a pregenerated library or created in an iterative manner through the application of transformations. In the latter case, structures are optimized against the QSAR fitness score by multiple rounds of mutation and selection, sometimes using a genetic algorithm [36]. However, the QSAR model cannot be expected to make accurate predictions on structures that are unrelated to the training set. For this reason a similarity or applicability filter is usually also applied to any structures under consideration. An
elegant approach, first used in 1993 by some of the pioneers [37] of inverse QSAR (then called inverse imaging) is to use graph-based structural fingerprints that are not only used as the QSAR descriptors but also for the construction of candidate chemical structures. A recent implementation of this approach by Martin [38] involves a complete enumeration of all possible alternatives within the neighborhood of the training set. This is a conceptually powerful approach because it is capable of producing locally optimal solutions. All iterative methods are nondeterministic and are based upon the premise that the possible solutions considered by a medicinal chemist will be even more limited in scope, if not quality.

5.3.5
**Multiple Objective Optimization**

As with medicinal chemistry in the real world, inverse QSAR involves trying to improve multiple properties of the molecules simultaneously. A common way of tackling this difficult problem is by the use of multiobjective optimization. Ekins et al. [39] demonstrated the use of Pareto multiobjective optimization, using transformation rules largely derived from the Drug Guru [30] system. One of the difficulties that they point out with such a system is that the multiple models used need to be reliable and predictive. This problem is addressed somewhat by the system developed by Segall et al. [40]. They employed a probabilistic scoring function that estimates the reliability of the predictions for each structure. The user is also able to control the relative influence of each of the component models.

In a groundbreaking paper Besnard et al. [41] used a process they called adaptive design to modify an existing drug’s polypharmacological profile. Some of the compounds they designed and synthesized not only had the desired “exquisite” selectivity and potency but were also active in vivo. Chemical structures were scored in multiple models of activity, synthetic tractability, novelty, and ADME properties. These were combined and a distance in multidimensional space to a single ideal achievement coordinate was calculated and optimized.

Inverse QSAR can produce some pretty strange looking structures. For instance, if the closeness to the training set is assessed using Tanimoto fingerprint similarity, because the presence or absence of a functional group is used rather than the count, a rash of a single functional group can result. For instance, Lewis [3] found multiple primary amines sprouting from aromatic rings in one of his test cases.

5.3.6
**Structure-Based De Novo Design**

The inverse QSAR methods described in Sections 5.3.4 and 5.3.5 could also be called de novo design. But this term is normally reserved for a stepwise construction of ligands in the presence of a 3D structure of the binding site [42]. In the presence of the receptor, individual fragments of the ligand can be optimized to achieve favorable interactions with atoms in the binding site. This is the main reason why
*de novo* design works by building up the ligand incrementally, optimizing each fragment, whereas inverse QSAR tends to look more at whole molecule properties.

The starting point for *de novo* design is a fragment or building block placed in the binding site. Its position and orientation can be obtained from X-ray crystallography or by using computational methods like docking or MCSS [43]. After a building block has been placed and optimized, the algorithm will select a new building block, attach it to the growing molecule using the connection rules, and optimize its conformation to achieve optimal interactions with the receptor. Then the next building block will be added and so forth. Building blocks can be single atoms or small fragments. Most of the early *de novo* programs used atoms as building blocks [44–47]. However, this fine-grain approach allows only a very limited chemical space to be searched and the resulting molecules can have low chance of being synthetically tractable. Most modern *de novo* design programs use fragments instead of single-atom building blocks. The use of selected fragments allows the biasing of constructed molecules toward desired molecular properties. For instance, drug-like molecules can be constructed from fragments of drugs, as in FOG or BOMB [48,49], or synthetically tractable molecules can be constructed from commercially available fragments as are used in SynSPROUT [50].

The rules, by which the fragments are connected, can also be used to drive the constructed compounds toward those that have desired properties. The FOG program [48] analyses the connection statistics between fragments in a training set of template molecules. These connection frequencies are then converted into probabilities, and used when the *de novo* algorithm decides which fragments to connect. The authors have shown that their algorithm was able to generate drug-like, natural product-like, or diversity-oriented compounds, depending upon the database used for training. Another molecular property that can be improved by the careful selection of the connection rules is synthetic accessibility. Some programs like TOPAS [51] and Flux [52,53] use Lewell’s RECAP rules [54] to first disconnect existing molecules and then recombine the resulting building blocks. This allows these programs to build realistic-looking molecules but still does not provide immediately applicable synthetic routes. This limitation is overcome by programs like DOGS [55] and SYNOPSIS [56] that use connection rules based on chemical reactions. As both the programs also exclusively use available building blocks, they make it very easy for the chemist to go from suggestion to synthesis. Connection rules are usually encoded in the *de novo* design program. A notable feature of the latest version of FTrees-FS [57,58] is that it is able to adapt its connection rules by dynamic programming to ensure that all or none of the generated molecules contain a required or forbidden functionality, respectively.

One of the obvious differences between *de novo* drug design and inverse QSAR approaches is that the former can use interactions with the receptor to predict relative activity and to exclude extensions that would clash with protein atoms. Structure-based methods can also employ receptor-based pharmacophores and hot spot prediction. We refer the reader to three very good reviews [59–61] for further details on *de novo* drug design. The authors of the SYNOPSIS system illustrate a combined approach using an in-house docking program. Their method included
an inverse QSAR component because they converted the interaction energy into a predicted \( pIC_{50} \), using a function trained on a set of known actives. This system is also noteworthy because it uses known reactions and reagents.

5.4 Virtual Synthesis

The construction of a virtual chemical compound can be done in variety of ways. Some of the construction methods illustrated in Figure 5.2 are mentioned earlier. In general they can be divided into those that are based on known reactions and are purely artificial. Virtual libraries sometimes serve as repositories of compounds or scaffolds to be considered for synthesis. In this case it is useful to have knowledge about their synthesizability and synthesis routes readily available. But how can synthesizability best be predicted?

5.4.1 Synthetic Tractability

The optimal method for evaluating the synthetic accessibility of a given compound is probably to search the chemical literature for cases where this or similar
molecules/scaffolds have been synthesized and to check the results with experienced organic chemists [62]. Unfortunately, this is not a feasible approach in the context of an automated algorithm to be used to filter the input to a large-scale virtual library. So computer based methods to evaluate synthetic accessibility have been developed.

Some of these methods are very detailed and require a lot of interaction with the user. They are aimed at supporting chemists in the actual planning of a synthetic route once they have decided which [63–67] specific molecule they want to make. Other methods are more focused on what is relevant in the context of this chapter: to apply a crude filter or ranking that prioritizes synthesizable molecules. This can be applied after the molecules have been generated, either by trying to fragment the suggested molecules according to computationally encoded retrosynthetic rules [54,68] or by comparing the scaffolds and molecules to molecules of known synthesizability [18,68,69]. An example of a de novo design package that works this way is SPROUT/CAESA [70].

5.4.2 Using Real-Life Reactions for in Silico Molecule Construction

Alternatively, instead of building up molecules agnostic of potential chemical reactions, and then trying to predict their synthesizability afterwards, another approach is to encode known chemical reactions in silico, and use them to build up the molecules.

Known chemical reactions can be found in the reaction databases like REAXYS [71]. More recently, we have observed a trend to not only capture and reuse reactions but also to analyze the frequency by which they are used in the pharmaceutical industry.

A highly cited analysis of “The medicinal chemist’s toolbox” was published by Roughley and Jordan in 2011 [72]. They collected over 7000 reactions from papers published by three major pharmaceutical companies in 2008. Their analysis of the usage frequency of the reaction types showed that heteroatom acylation and alkylation accounted for nearly 50% of all reactions, and that medicinal chemists relied heavily on a relatively small number of reaction types in general. This analysis supports the idea that by encoding a manageable number of reactions, one could reasonably mirror what medicinal chemists do, and cover a sufficient percentage of the usually accessed chemical space. This draws our interest to the collections of common reactions in the literature, some of which have been assembled to be used in a de novo design context.

A notable study is the one from Schürer et al. [73], who published a collection of 54 reactions. The supporting information contains the names and diagrams of the reactions, but they are not in a computer-interpretable format. The first publication that encoded a similar collection of reactions in computer readable format was from Vinkers et al. [56]. The 75 reactions underlying their de novo tool SYNOPSIS are available from the authors as Reaction SMARTS [74]. More
recently, Hartenfeller et al. [75] assembled a collection of 58 reaction schemes for their DOGS program [55]. Half the reactions in their data set are ring formations. This type of reaction was found to be especially useful for the generation of fundamentally new chemotypes or scaffolds. In comparison, the Schürer collection only included 14 that were ring-closing reactions [73]. The SMARTS of the reactions compiled by Hartenfeller et al. can be downloaded free of charge from the publisher’s website. In a later publication [76], the authors showed that about 10% of all analyzed bioactive molecules and drugs could be constructed using this set of 58 reactions. This analysis gives credence to the concept of “using a limited number of reactions to rebuild a lot of relevant medicinal chemistry space.”

5.4.3

**Readily Synthesizable Compounds**

Computer encoded chemical reaction can be used to ensure that all compounds in a virtual library are synthesizable. But this can also be taken a step further by creating a virtual library of every molecule that could be made using a validated reaction. If an organization uses encoded versions of all its easily applicable chemical reactions and building blocks easily accessible in-house or commercially to build up a database, they can get a repository of everything they could make quickly. If a technology exists for searching and filtering this huge combinatorial chemical space, it can be used to quickly take decisions on “which compound should be made and tested next?” The following four notable examples of this approach will be described below: Tripos’ AllChem [77], Boehringer Ingelheim’s (BI) BI CLAIM [58], Pfizer’s PGVL [78,79], and AZ Virtual Library (AZ-VL) [80].

5.4.3.1 **Construction**

Constructing and searching very large virtual chemical spaces poses technological challenges that have been addressed differently by these four companies. The current technological limit for storing enumerated chemical compounds in a searchable format is around $10^8$ compounds [10,78,81].

Tripos’ AllChem started with 7000 reagents. Cramer et al. recursively applied a set of 100 carefully encoded reactions from their experience and the literature to these reagents. This process resulted in five million theoretical building blocks, called synthons. The $10^{20}$ molecules that these synthons could be combined into were not explicitly enumerated. Instead, AllChem stored the synthons and the rules to connect them.

Pfizer and AZ bypassed the need for complete enumeration by storing only the reaction schemes and associated reactants in their PGVL/AZ-VL. Only during searches does PGVL enumerate small subsections of the complete Virtual Library space (which would be $10^{14}$–$10^{18}$ in size). Many of the synthetic protocols they used (PGVL: 1244 and AZ-VL: 4000) to build up their virtual libraries were manually encoded and validated by chemists from their
combinatorial chemistry groups. In addition, AZ revealed in the paper how an automated algorithm had been used to extract and encode additional reaction schemes from the electronic laboratory notebooks (ELNs), adding 20,000 libraries to the 4000 hand curated ones.

To avoid having to store $10^{11}$ molecules, BI built up its library BI CLAIM by using the “fragment space” approach [82], storing only fragments and linkage rules. To stay clear of one of the major problems with fragment spaces, the lack of synthesizability [54,73], Lessel et al. leveraged Boehringer’s in-house knowledge of feasible chemistry for the construction of their BI CLAIM database.

### 5.4.3.2 Searching

Standard similarity search algorithms based upon molecular fingerprints [83,84] are limited to explicitly enumerated libraries of up to $10^8$ molecules [57]. So systems like these huge virtual libraries also had to find new ways to search their large non-enumerated spaces. Tripos used their proprietary Topomer similarity search technology [85,86], which allowed searches of AllChem within a few hours. BI CLAIM used FTrees-FS [82]. It compared the feature trees of the fragments in the query with fragments in the database. Out of the fragments with the highest similarity, it dynamically created molecules with high similarity to the query.

AstraZeneca’s virtual library was searched in basis product space [87,88]. In a second stage, a mathematical combination of the binary fingerprints of the most similar basis products was used to quickly approximate a fingerprint for the enumerated compounds. The similarity of this approximated fingerprint to the query is then used to rank the output. The authors demonstrated that this considerably faster similarity comparison still returned hits of equivalent quality to the results from slower exhaustive screens.

The PGVL methodology uses two different search techniques, LEAP1 or LEAP2, depending on whether the query molecule can be disassembled using combinations of virtual reactions from PGVL or not. If disconnection is possible, the reactants are used in similarity queries against the reactants in the corresponding pools in PVGL. Wherever hits for similar reactants are found, the relevant subsets are then explicitly enumerated and searched using standard fingerprint-based similarity. Query molecules that do not lend themselves to LEAP1 are used with the alternative search method LEAP2. This method works in a similar way to the basis product approach applied by AZ (see above). One difference is that it applies an asymmetric similarity measure [89] to avoid some of the problems that arise from similarity comparisons in basis product space. Another difference between LEAP2 and the method used by AZ is that the latter calculates approximated fingerprints of final molecules, whereas the former explicitly enumerates areas of PGVL, which have a higher probability of containing molecules similar to the query. These areas are then fully enumerated and searched using standard fingerprint or molecular keys similarity.
5.4.3.3 Outside Big Pharma
A similar concept, but on much smaller scale, was realized with the noncommercial TIN database [90]. Dorschner et al. used 12 reaction schemes to build a virtual library of over 28 million compounds. Because of the moderate size, this library could still be fully enumerated and searched in that format. The authors actually made the structures available online.

A phenomenon observed by the authors is that some commercial chemical suppliers list virtual compounds in their catalogs. These compounds typically have lead times of several weeks—enough to initiate the synthesis only when customers order the compound. Examples of these virtual libraries are the “make-on-demand” category in ZINC [11] and the Enamine’s “REAL” database of $2 \cdot 10^7$ compounds [91].

5.4.4 Iterative Approaches

Another way to combine virtual libraries and known reactions is exemplified by the two de novo design programs DOGS [55] and SYNOPSIS [56]. Both programs use sets of known reactions to assemble virtual molecules. DOGS’ aim is to find bioisosteric replacements for known active ligands. So it directs the growth of the virtual molecules on maximizing the ligand similarity to a given query. SYNOPSIS does not have a fixed fitness function that enables its use for a wider range of applications. Vinkers et al. show in their paper how either a dipole calculation or a protein–ligand affinity as predicted by molecular docking can be used as a fitness function.

5.5 Visualizations of Scaffold Space

Various ways of visually assessing the diversity and frequency of scaffolds, particularly ring systems in a compound collection, have appeared in literature. Some like bubble trees [92] and scaffold cloud [93] spread pictures of the scaffolds out in 2D. The size of each picture is proportional to scaffold frequency. Other techniques position similar structures close to each other in space using dimension reduction techniques. The most commonly used technique is principal components analysis (PCA), for example [40]. Kohonen maps have also been employed [17]. A simple but appealing alternative is to plot the geometric distance between substitution points against the angle between R group vectors [94]. These latter techniques provide a visual way of assessing how a new scaffold fits within an existing collection. This sort of approach has been used for “hole filling” in compound collections. Given what we know about the size of chemical space, this term now seems naively optimistic.

Several approaches extend the idea of Bemis and Murcko frameworks [7] by breaking compounds into a hierarchy [95] or network [96] of ring-based scaffolds.
In Scaffold Hunter [6], the ring assemblies of a molecule are simplified ring by ring in a consistent manner until a monocyclic root is reached. This process creates a tree structure of sequentially smaller substructures as nodes. When a series of compounds is analyzed in this way, differences in scaffolds appear as branches in these trees. Where the scaffolds of compounds are connected by a common node, the presence of a common subscaffold is revealed. Each node that matches one or more complete structure can be annotated with biological activity or other properties. The authors hypothesize that, where active compounds are connected in a tree, compounds built around the virtual scaffolds that connect them are also likely to be active. They tested this hypothesis by testing compounds containing virtual scaffolds identified in this way and achieved an 8% hit rate.

5.6 A Perspective on the Past and the Future

The freedom from the restrictions of physical chemistry has led to many innovative explorations of virtual chemical space. These range from the enumeration and analysis of complete areas of chemical space to algorithms that hunt through chemical space in a stepwise manner. These studies continue to reveal what is special about the molecules that chemists have chosen to make and have proven to be biologically active. One aim is to make virtual molecules more appealing to medicinal chemists. Another is to improve the quality of compounds they make in the future. These aims can be achieved not only by learning from past successes and failures of the finished products but also by capturing and exploiting knowledge of how they were constructed. However there is still much to be learnt. Synthetic medicinal chemists will perhaps be relieved to learn that computers are still some way from equaling the ability of human beings to design drugs.

With the advent of freely accessible databases of SAR data such as within ChEMBL [8], the patterns, successes, and failures of past chemistry programs are there to be discovered. Databases of known reactions are not quite at this level of accessibility but the widespread use of electronic lab notebooks is at least providing computational access to in-house syntheses. Unfortunately, commercial databases of literature reactions, although invaluable to chemists, are designed to be accessed manually and do not usually allow programmatic queries. The paucity of freely downloadable resources has probably held back the development of virtual chemical library exploration by decades.

One of the challenges that researchers in pharmaceutical companies wishing to access all readily synthesizable compounds have faced is the sheer number of such structures. Creative ways have been found to help deal with this problem (see Section 5.4.3), but the field would surely benefit from more academic input and community-wide resources. Systems created by big pharma, such as BI CLAIM, PGVL, and AZ-VL are attractive when the main starting information in a project is
the 2D structure of one or more active ligands. In these cases 2D similarity methods are the most promising virtual-screening methods. And these could really benefit from the knowledge about synthetic accessibility, contained in these systems. As far as we know, these types of systems, where very large virtual libraries are scanned, have yet to be coupled to structure-based virtual screening or de novo design or exploited for scaffold hopping.

While the exploration of virtual chemical space is widespread in the literature, very few authors have focussed their efforts on chemical scaffolds explicitly. However, since the term scaffold is often used in a retrospective sense to highlight one or more common substructures within a congenic series, perhaps this is not surprising. It would be interesting to see whether virtual series of compounds also contain these common central cores. Do conserved scaffolds result from pragmatic behaviors of medicinal chemists, such as the exploration of one position of variation at a time using a carefully prepared starting material and chemistry amenable to parallel synthesis? The tendency to follow this path may be compounded by the usual practice of basing chemical patents around Markush substructures. If virtual libraries are built using inverse QSAR or de novo design, without these considerations built into the algorithm, do other patterns emerge? Do virtual libraries, built using virtual synthesis, result in different patterns from those built using synthetic tractability scores? These questions remain to be answered. Ultimately, what we want to know is whether computational algorithms can help improve the ways in which chemists search chemical space in the real world.

At the time of writing, the closest that researchers had got to exploring virtual scaffold space directly was the complete enumeration of small fragments and aromatic ring systems. Scaffolds are by definition smaller than whole molecules, and size of chemical space decreases exponentially with the number of atoms. So scaffold space is more accessible to comprehensive studies. Ring systems are perhaps the best place to start as they are often used as scaffolds. The methodology used for the complete enumeration of aromatic rings systems can just as easily be applied to other small scaffold subsets. Furthermore, small scaffolds of interest are likely to be contained in the GDB-13 or the GDB-17. Unfortunately, this latter database is not offered for download in its entirety, presumably because it is too large.

We should not ignore the successful drug discovery uses to which virtual chemical space exploration has already been put. Those resulting from de novo design are reviewed elsewhere [59,60], so they will not be listed here. The successful design of molecules with particular activity profiles using multiple objective inverse QSAR [41], which was published during the writing of this chapter, should alone provide enough encouragement to others to think of trying this approach themselves. However, the use of simpler techniques such as producing small virtual libraries that contain all the ideas that chemists and modelers can come up with, followed by selection using property calculations and activity predictions, is already commonplace. One difficulty that computational chemists have had is that PhD-level organic chemists like to make compounds that they or their immediate peers have thought of. Providing tools that are under the control of these chemists is one way to achieve more widespread adoption. Drug
Guru [30] and the Molecule Evoluator [31] are examples of this sort of tool. With the rise of contract chemistry, where drug designers farm out syntheses to synthetic chemists in remote laboratories, these sorts of tools may prove invaluable.

Developments that might enrich the exploration of virtual chemical space in the future are the treatment of starting materials and products as nodes of a network. This is the approach taken by Grzybowski and coworkers [97] for real chemistry. The benefits of this approach include a unique view on the behavioral patterns in published chemistry over the years, the identification of important compounds as highly connected hubs in the network [98], and the identification of more optimal routes to particular products as different paths through the network [99]. It does not take a great deal of imagination to see how this methodology could be expanded to include virtual synthesis.

Another development that will benefit from the work described in this chapter is that of automated chemistry. When robots are made to carry out reactions, readily synthesizable compounds become readily programmable compounds. Eli Lilly has created an automated chemistry laboratory on which chemists around the world can order compounds from their sites [100]. The next obvious step is to allow computers to design the compounds. Cyclofluidic [101] is a company set up to use flow devices for this purpose. Eve is a robotic system being developed to carry out drug discovery automatically [102]. It seems likely that methods for exploring virtual chemical space will be put to the test experimentally, without human guidance (or interference, depending on your point of view), in the near future. This will highlight the benefits and limitations of current approaches and should help the field to progress further.

References


27 Papadatos, G., Alkarouri, M., Gillet, V.J., Willett, P., Kadirkamanathan, V., Luscombe, C.N., Bravi, G., Richmond, N.


References


101 Cyclofluidic Ltd., http://www.cyclofluidic.co.uk (29 July 2013).
Part Two
Scaffold-Hopping Methods
6
Similarity-Based Scaffold Hopping Using 2D Fingerprints

Peter Willett

6.1 Fingerprints

Scaffold hopping is one specific form of the more general task of virtual screening, that is, the identification of those molecules in a chemical database with high a priori probabilities of activity against some biological target. Here, we focus on similarity-based virtual screening, where the database is searched for those molecules that are most similar to a reference structure, which is a molecule that is already known to exhibit the activity of interest. In similarity-based scaffold hopping, the aim is to identify those database structures that have high levels of overall similarity to the reference structure but that contain different ring systems.

Two factors are of importance in the calculation of intermolecular similarity. The first factor is the type of representation that is used to characterize the reference and database structures [1–3]. There are three basic types [4]: sets of computed molecular properties (e.g., molecular weight, polar surface area, and numbers of heteroatoms), 2D representations encoding molecular topology, and 3D representations encoding molecular geometry. This chapter considers only the second class of representations, specifically 2D fingerprints; other types of topological representation that can be used for the calculation of similarity include topological indices and chemical graphs (both complete and reduced) [5]. A 2D fingerprint is normally a binary vector in which the individual bits encode the presence or absence of a chemical substructure (or fragment) in a molecule (see below). The second factor is the similarity coefficient that is used to quantify the degree of resemblance between the reference-structure and database-structure representations, that is, between two fingerprints in the present context. There are many coefficients that can be used for this purpose, such as the Tanimoto coefficient that has been shown in comparative studies to be well suited to a wide range of applications in chemoinformatics [3,6,7]. Given two fingerprints containing \(a\) and \(b\) nonzero bits, \(c\) of which are in common, the Tanimoto coefficient is defined to be \(c/(a + b - c)\), with the extremal values of zero or unity denoting fingerprints having no bits or all bits in common (and hence denoting that the corresponding molecules have no or all fragments in common, respectively). Most studies of scaffold hopping have employed the
Tanimoto coefficient, but others have also been used for this purpose, such as the related Tversky coefficient [8].

There are two ways in which the presence of 2D substructures can be recorded in a fingerprint [4]. The better-established, and conceptually simpler, dictionary-based approach requires a list of known substructures with, normally, each of these being associated with a specific bit in the fingerprint, so that there is a one-to-one relationship between the presence of a substructure in a molecule and the setting of a bit at a specific location in the binary vector. The alternative molecule-based approach differs in two ways from the dictionary-based approach. First, the fingerprint encodes many-to-many relationships, with hashing algorithms being used to set several bits to denote the presence of a particular substructure in a molecule and with the inevitable collisions from the hashing, meaning that several, or many, different substructures are associated with each bit. Second, there is no predefined dictionary of preselected substructures; instead, a class of substructure is chosen, for example, a chain of four atoms and the interconnecting three bonds, and then each and every example of this substructural class in a molecule is submitted to the hashing algorithms. Well-known examples of the dictionary-based and molecule-based approaches are MDL key fingerprints (from Accelrys Inc.; available at http://www.accelrys.com) and Daylight fingerprints (from Daylight Chemical Information Systems Inc.; available at http://www.daylight.com), respectively; the Unity fingerprint (from Tripos Inc.; available at http://www.tripos.com) is largely based on the molecule approach but also has some bits that are set using the dictionary approach.

Fingerprints were originally developed to enhance the efficiency of chemical substructure searching, that is, the ability to identify all of the structures in a database that contains a user-defined partial structure (such as a penicillin ring system in a search for potential antibiotic molecules). However, it was soon realized that fingerprints could additionally be used to calculate the degree of resemblance between two complete structures. The resulting similarity values could be computed extremely rapidly and were found to correspond well with chemists’ subjective views of topological similarity, with the result that similarity-based virtual screening rapidly established itself in lead discovery programs in the pharmaceutical and agrochemical industries [2,9].

Thus far, we have considered fingerprints as encoding only the presence or absence of a substructure (as denoted by a bit (or bits) being set to 1 or 0, respectively). There is, however, no reason why the binary vector cannot be replaced by an integer or real vector in which the individual elements of the fingerprint store weights, reflecting the relative importance of substructures. Then a substructure with a high weight (e.g., a substructure that occurred not once but several times) in both the reference structure and a database structure would make a larger contribution to the computed similarity than a substructure that was common to both but that had a lower weight. The CATS fingerprints discussed in Sections 6.2 and 6.3 in the chapter make use of weighting.

The effectiveness of 2D fingerprints for scaffold hopping can be tested using both retrospective and predictive studies. Retrospective studies are based on the use of test data sets for which sets of actives and inactives (or, more usually, presumed
inactives) are known, thus allowing detailed benchmarking studies that compare
the screening effectiveness of different fingerprint methods. Predictive studies,
conversely, describe real-world applications in which operational implementations
of fingerprint-based virtual screening are used to identify novel bioactive
molecules. These approaches are described in Sections 6.2 and 6.3 of this chapter,
before we present our final conclusions.

6.2 Retrospective Studies of Scaffold Hopping Using 2D Fingerprints

Since the first reports of fingerprint-based similarity methods over two decades ago
[9,10], a very large literature has developed, reporting the evaluation and
comparison of such methods [4,11,12]. It was not, however, till 1999 that Schneider
et al. provided probably the earliest report of using 2D fingerprints for scaffold
hopping [13]; several retrospective studies have now been reported [14–18].

These studies will be illustrated by presenting some of the results from a recent
comparison by Gardiner et al. [18] of the effectiveness of a range of types of
fingerprint, when used for scaffold hopping on three standard test databases: the
accelrys.com), the World of Molecular Bioactivity (WOMBAT) (from Sunset
Molecular Discovery LLC; available at http://www.sunsetmolecular.com), and the
Maximum Unbiased Validation (MUV) (from the Carolo-Wilhelmina Technical
University in Braunschweig; available at http://www.pharmchem.tu-bs.de/lehre/
baumann/MUV.html) databases. In what follows, we shall focus on just the results
obtained by Gardiner et al. with the MDDR data set, which contained 102,516
molecules and had 11 sets of active molecules for which scaffold-hopping searches
were carried out. Brief details of these classes are listed in Table 6.1. In like vein, we

<table>
<thead>
<tr>
<th>Activity class</th>
<th>Active molecules</th>
<th>Active scaffolds</th>
<th>Mean similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5HT1A agonists (5HT1A)</td>
<td>827</td>
<td>271</td>
<td>0.34</td>
</tr>
<tr>
<td>5HT3 antagonists (5HT3)</td>
<td>752</td>
<td>237</td>
<td>0.35</td>
</tr>
<tr>
<td>5HT reuptake inhibitors (5HT)</td>
<td>359</td>
<td>126</td>
<td>0.35</td>
</tr>
<tr>
<td>Angiotensin II AT1 antagonists (AT1)</td>
<td>943</td>
<td>285</td>
<td>0.40</td>
</tr>
<tr>
<td>Cyclooxygenase inhibitors (COX)</td>
<td>636</td>
<td>139</td>
<td>0.27</td>
</tr>
<tr>
<td>D2 antagonists (D2)</td>
<td>395</td>
<td>187</td>
<td>0.35</td>
</tr>
<tr>
<td>HIV protease inhibitors (HIV)</td>
<td>750</td>
<td>331</td>
<td>0.45</td>
</tr>
<tr>
<td>Protein kinase C inhibitors (PKC)</td>
<td>453</td>
<td>134</td>
<td>0.32</td>
</tr>
<tr>
<td>Renin inhibitors (REN)</td>
<td>1125</td>
<td>339</td>
<td>0.57</td>
</tr>
<tr>
<td>Substance P antagonists (SUB)</td>
<td>1246</td>
<td>380</td>
<td>0.40</td>
</tr>
<tr>
<td>Thrombin inhibitors (THR)</td>
<td>803</td>
<td>295</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Each row of the table contains the type of active molecule in a class, with an abbreviation in brackets;
the number of actives of the class in the entire 102,516 molecule data set; the number of active
scaffolds (see text); and the mean pairwise similarity, calculated using the Tripos Unity fingerprint and
averaged over all pairs of active molecules in the class.
shall focus on just one of the six different types of fingerprints that were studied: this was the ECFP4 (where ECFP denotes extended-connectivity fingerprint) fingerprint that is provided in the Accelrys’ Pipeline Pilot software [19]. ECFP4 encodes circular substructures of diameter four bonds centered on each of the heavy atoms in a molecule, with the substructural occurrences hashed into a fixed-length fingerprint containing 1024 bits. The other fingerprints studied were the Accelrys FCFP4, Tripos Unity, BCI, Daylight, and MDL fingerprints [18].

Given a reference structure belonging to a specific activity class, the reference structure was matched against each of the database structures in turn, the resulting Tanimoto similarity values sorted into decreasing order, and then a cutoff applied to retrieve the top 1% of the ranked database. This was carried out using 10 different reference structures for each of the activity classes. The scaffold-hopping capability was quantified by the numbers of distinct active scaffolds retrieved in a search, where an active scaffold was a scaffold that occurred in one of the active molecules for a class and where the definition of a scaffold was that provided by the Murcko scaffold routine in the Pipeline Pilot software. Three different performance criteria were adopted: criterion A was the percentage of the active scaffolds identified in all the molecules (not just the active molecules) in the top 1% of the ranking, criterion B was the corresponding percentage for just the top-ranked active molecules, and criterion C was the percentage of the top-ranked active molecules having a scaffold different from that of the reference structure. The screening performance for each activity class was obtained by calculating first the appropriate criterion value for each individual search and then the arithmetic mean averaged over the 10 chosen reference structures for that class.

Figure 6.1 summarizes the results obtained in the ECFP4 searches of the MDDR data set, showing the calculated values for the three performance criteria for each of the 11 activity classes: for example, in the 5HT1A searches, values of 21.0, 12.8, and
10.1% were obtained using criteria A–C. Inspection of the figure makes very clear
that the ECFP4 fingerprints enable scaffold-hopping searches of chemical
databases to be carried out, which are far more effective than random screening
(which, e.g., would be expected to retrieve just 1% of the active scaffolds using
criterion A). Inspection of the full results in the original publication shows that
such enrichments over random screening were obtained with all the fingerprint
types in all of the searches that were carried out (a total of 38 activity classes across
the three data sets), with the sole exception of some of the searches for SF1
inhibitors in the MUV data set [18]. It might be expected that the structural
diversity of an activity class would affect the level of enrichment that is obtainable,
and the best results here were indeed obtained with the relatively homogeneous set
of renin inhibitors (as denoted by the mean similarity of 0.57 in Table 6.1), but the
results obtained were far from unequivocal with, for example, the notably poorer
performance for the HIV class despite it having the second highest mean similarity
in Table 6.1. That said, the MDDR and WOMBAT results in general were markedly
higher than those for the MUV data set, which was specifically designed to involve
sets of highly diverse actives in each class.

Further support for the view that 2D fingerprints can indeed be used in scaffold-
hopping searches comes from analogous retrospective studies by Vogt et al. [14]
and by Renner and Schneider [17].

Vogt et al. used carefully selected sets of known actives for 17 biological targets,
adding these to about 500,000 presumed inactives from the ZINC and ChEMBL
.uk/chembl/, respectively) and then carrying out scaffold-hopping searches that
were evaluated in a manner similar to criterion B mentioned earlier. The
experiments involved five different types of fingerprint (including ECFP4) and data
fusion [20] searches in which not one but five reference structures sharing a
common scaffold were used, with each database structure being scored on the basis
of the largest of the five resulting similarities. Averaging across the five fingerprints
and 17 activity classes, they concluded that retrieving the top 0.5–1% of a ranked
database would result in the identification of about 25% of the available active
scaffolds (although notably greater performance levels could be expected in some
cases).

Both Gardiner et al. and Vogt et al. included MDL keys [21] among the
fingerprints that were compared, and these also figured in the comparison reported
by Renner and Schneider [17]. This study used the CATS 2D autocorrelation
function, a fingerprint in which the encoded substructures describe pairs of atoms
(each characterized by their pharmacophore type) and the numbers of bonds
separating them. Searches were conducted for 10 activity classes in 4075 molecules
from the COBRA (Collection of Bioactive Reference Analogues) database [22];
successful scaffold hopping was again observed, with little difference in the
performance of the CATS and MDL fingerprints (and also of other versions of
CATS that described 3D molecular shapes or surfaces). Like Gardiner et al., Renner
and Schneider noted the effect of class diversity, suggesting that the MDL keys,
which had originally been developed to support 2D substructure searching, might
be most appropriate for scaffold hopping when structurally homogeneous sets of actives were being sought.

6.3 Predictive Studies of Scaffold Hopping Using 2D Fingerprints

Retrospective studies such as those in Section 6.2 allow detailed comparisons to be made of the effectiveness of different types of fingerprints (or of similarity coefficient or of retrieval strategy) and demonstrate the clear potential of 2D fingerprints for scaffold hopping; whether this potential can be realized in practice is discussed below, where we summarize the results of several predictive studies of virtual screening.

Shoda et al. describe the successful use of fingerprint-based virtual screening in a study of structurally diverse growth hormone secretagogue (GHS) agonists [23]. This is an important therapeutic area since GHS is involved in the release of growth hormone in the pituitary gland. GHS agonists may hence provide a way of treating growth failure, which can cause adults to be affected by low muscle strength and high levels of cholesterol, and may also have application in the treatment of osteoporosis and obesity *inter alia*. The starting points for the investigation were two previously patented compounds (I and II) that had been reported as showing GHS activity. Similarity searches of an in-house database identified 58 molecules that were similar to I or II after filtering the top-ranked molecules using Lipinski rules and lists of pharmacophores and reactive groups. The fingerprints used in these searches were MFP2 fingerprints that encode three numeric property values (the numbers of aromatic bonds, hydrogen bond acceptors, and the fraction of rotatable bonds) and 40 substructural keys (describing specific atoms types, functional groups, and bond patterns) [24]. The 58 hits were tested for GHS activity in a calcium flux assay, with the most active molecule being III below. This formed the basis for an analog study in which 117 molecules from the in-house database, which were similar to III, were tested in the assay. The 14 that showed activity were subsequently tested in a radioligand binding assay, and 13 of these were found to bind with $K_i$ values in the range $0.22$–$2.0 \mu M$. 

![Chemical structures](image)
In a second successful example, Wang et al. describe the identification of novel agents for the treatment of advanced melanoma that is the most deadly form of skin cancer and is only poorly treated by the agents that are currently available [25]. A previous study had identified a new class of thiazole analogs that were highly potent against *in vitro* melanoma cells, and one of these (shown in IV below) formed the basis for a similarity-based virtual screening study in which it was initially searched against a file of about 350,000 structures using five different extended-connectivity fingerprints (including the ECFP4 fingerprints mentioned previously). The top 400 database structures from each of these searches were ranked in decreasing similarity order and then the mean rank calculated for each of the resulting distinct structures, when averaged over the five separate sets of top-ranked structures, providing a further example of the potential of data fusion for scaffold hopping [20]. Of the top-ranked structures, 330 were then tested for activity against two cancer cell lines (A375 and B16-F1): 8 of these proved to be active, the most potent being V, which had an IC$_{50}$ value less than 1 μM on the A375 cells and good selectivity between cancer and normal cells. This hit is clearly very different in structure from IV, and Wang et al. noted that the other actives identified in the fingerprint searches also demonstrated a marked level of scaffold hopping, with, for example, one of the actives having the trimethoxyphenyl ring in IV replaced by an isothiourea group.

The final example is described by Franke et al. in a study of inhibitors of 5-lipoxygenase (5-LO), a potentially important drug target that is involved in the mediation of inflammatory and allergic reactions [26]. Of the known 5-LO
inhibitors 43 were selected from the COBRA database (see above) and matched against the MEGx database of natural products and the NATx database of semisynthetic small molecules derived from natural products (from AnalytiCon Discovery; available at http://www.ac-discovery.com). The 10 most similar hits for each of the known inhibitors were retrieved and 18 of these selected for screening for inhibition of leukotriene synthesis in both intact cells and a cell-free assay, resulting in the identification of two novel, submicromolar inhibitors (VI and VII). These were then screened against a large NATx combinatorial library built around scaffolds derived from α-santonin, a common sesquiterpenelactone occurring in different Artemisia species. Whereas the initial similarity searches had used CATS fingerprints, the second set involved not only CATS itself but also MDL keys and two versions of CATS that encoded atomic charges. The four similarity searches retrieved a total of 17 analogs of VI and VII; these had comparable activities to the reference structures, with a subsequent SAR study, suggesting that VI and its analogs operate by a different mechanism from VII and its analogs.

Further examples of successful scaffold hopping include studies by Peukert et al. of novel blockers of the voltage-gated potassium channel Kv1.5 using Tripos Unity fingerprints [27] and by Noeske et al. of selective allosteric mGluR1 antagonists using CATS fingerprints [28]. In both cases, the 2D fingerprint search provided an initial hit for synthesis and SAR studies that eventually resulted in a novel, submicromolar lead compound.

6.4 Conclusions

Many different techniques have been described in the literature for virtual screening, but similarity searching continues to be extensively used in the early stages of lead discovery programs [4,11,12]. There are several reasons for this: it can be used when little or no information may be available as to the nature of the pharmacophore and/or the binding site (indeed, a single active reference structure suffices to act as a starting point); similarity searching is inherently very simple and
there is a wide range of software available for this purpose; it has been demonstrated to be successful in both retrospective and predictive analyses.

The examples of scaffold hopping described in Sections 6.2 and 6.3 have used various types of 2D fingerprint for the computation of similarity, and it would hence seem reasonable to assume that more sophisticated similarity measures that make use of 3D structural information would provide a still greater scaffold-hopping capability. This would indeed appear to be the case. For example, Tresadern et al. studied a range of ligand-based virtual screening methods in both retrospective and prospective studies of corticotrophin-releasing factor 1 antagonists, and found that 3D similarity methods (based on shape and on electrostatic charge) were better for scaffold hopping [29]; and Moffat et al. came to the same conclusion in a retrospective study of some of the MDDR activity classes described in Section 6.2 [30]. Other studies, however, have reported 2D similarity, searching outperforming 3D methods in terms of the overall numbers of actives retrieved [31–33], suggesting that scaffold hopping can be maximized by combining both sorts of structural information [16,34,35].

Finally, it should be noted that this chapter has focused solely on the use of database searching for fingerprint-based scaffold hopping. 2D fingerprints can also be used in other ways for this purpose, for example, by adopting a de novo approach in which potential actives are designed (rather than searched for) on the basis of similarity to known actives [36,37].

References


7
CATS for Scaffold Hopping in Medicinal Chemistry

Christian P. Koch, Michael Reutlinger, Nickolay Todoroff, Petra Schneider, and Gisbert Schneider

7.1
Chemically Advanced Template Search

Vector representations of molecular structures are commonly and successfully used in medicinal chemistry for virtual screening and compound design. Typically, the numerical values of such a molecular representation (so-called “descriptor values”) are computed by a function $y = f(x)$, where $x$ denotes a molecular structure and $y$ the corresponding vector representation. While the mapping $x \rightarrow y$ is uniquely defined, the back projection $y \rightarrow x$ usually is ambiguous. In other words, mapping from the instance $y$ to the set of compound structures $X$ is equivalent to finding the associated subset $U = \{x \in X | y = f(x)\}$ as the archetype of $y$ (Figure 7.1). In fact, this property of descriptor functions is necessary for scaffold hopping, as only a molecular representation that considers various structural elements like ring systems and certain functional groups as equivalent will be useful for finding compounds that share a desired property or activity (functional similarity) but differ in their constitution and configuration (structural dissimilarity). One might therefore state that molecular representations need to be “fuzzy” (permissive) for the purpose of scaffold hopping [1,2].

The concept of fuzzy molecular representations has been applied to chemical similarity searching since the beginnings of computational medicinal chemistry. In this context, Schneider et al. coined the term “scaffold hopping” in a publication dating back to 1999 that introduced chemically advanced template search (CATS) as a variation of similarity searching using topological pharmacophore features [3]. CATS molecular representations may be considered as a variation of Carhart’s atom pair descriptors [4], and a play on the acronym “ATS” that stands for Autocorrelation of Topological Structure (Eq. (7.1)) [5,6]:

$$\text{ATS}_d = \sum_{i=1}^{A} \sum_{j=1}^{A} \delta_{ij} \cdot (w_i \cdot w_j),$$

(7.1)
where $d$ is the considered topological distance between atoms $i$ and $j$ (expressed as numbers of bonds along the shortest path between graph vertices $i$ and $j$), $A$ is the number of atoms, $w$ is an atomic property (electronegativity, partial charge, etc.), and $\delta$ the Kronecker delta that evaluates to one if vertices $i$ and $j$ are spaced $d$ bonds apart, zero otherwise. The distance range considered has critical influence on the ambiguity or permissiveness of these types of correlation-vector representations. Limiting the range for $d$ to, say, 1–4 bonds only captures local molecular property patterns, while allowing distances of, for example, 1–10 bonds captures global property arrangements, which is related to the concept of radial fingerprints. The concept of autocorrelation and cross-correlation of molecular properties was pioneered by Gasteiger and coworkers in the 1990s and extended to three-dimensional (3D) conformations [7,8]. CATS modifies the ATS descriptor by the following:

i) using binary attributes (potential pharmacophore points (ppp)) instead of real-valued atomic properties, and

ii) introducing a scaling factor $\lambda^{-1}$ (eq. (7.2)):

$$\text{CATS}_{d}^{F_1,F_2} = \frac{1}{\lambda} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \delta_{ij} \cdot (F_{1i} \cdot F_{2j}),$$  \hspace{1cm} (7.2)$$

where vertices $i$ and $j$ must possess the pharmacophore features $F_1$ and $F_2$, respectively, to compute a value other than zero. Simply put, one counts the number of atom pairs spaced $d$ bonds apart, which possess the respective pharmacophoric feature (Table 7.1). Most importantly, for each feature pair $F_1$ and $F_2$, the resulting raw count should be divided by $\lambda$, which is the total number of atoms possessing feature types $F_1$ and $F_2$. This term reduces the influence of the molecular graph size on the CATS descriptor and dominance of individual features (a concept similar to using Shannon entropy for descriptor scaling [9]). As an example, let us consider compound 1 (fluoxetine, Prozac™), which is first converted to its molecular graph 2. Then feature types are assigned to the vertices
### Table 7.1 CATS2 atom types.

<table>
<thead>
<tr>
<th></th>
<th>Carbon, lipophilic</th>
<th>Halogens</th>
<th>Nitrogens</th>
<th>Oxygen</th>
<th>Acids</th>
<th>Hydroxy</th>
<th>Charges</th>
<th>Aromatic</th>
<th>CDK type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Cl</td>
<td>Br</td>
<td>I</td>
<td>[N,H1]</td>
<td>[N,H2]</td>
<td>[N,H4]</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Lipophilic</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donor</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aromatic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

a) Carbon atoms adjacent to heteroatoms, but not halogens, receive no pharmacophore feature.
b) In CATS2, an atom is either lipophilic or aromatic. CDK: Chemistry Development Kit [10].
The molecular structure (1) is reduced to the molecular graph (2), and feature types are assigned, where R: aromatic, L: lipophilic, A: hydrogen bond acceptor, and D: hydrogen bond donor (3). In step 3, atom pairs for all feature pairs are counted. The final descriptor values must be divided by the corresponding $\lambda$ values for proper scaling (step 4). Note that not all vertices are considered “pharmacophoric.”

An online software tool that implements CATS is available on the World Wide Web at http://modlab-cadd.ethz.ch/.

7.2 Retrospective Evaluation of Enrichment and Scaffold Hopping Potential

Despite its primary purpose of providing a sufficiently abstract molecular representation that enables molecular scaffold hopping, several retrospective benchmark studies assessed the ability of CATS to retrieve known ligands of a macromolecular target from a large screening compound pool [11,12]. Of note, in none of these analyses or implementations CATS descriptors were properly scaled [13]. Here, we present a comparative study to analyze the influence of parameter settings on the retrieval of active compounds (enrichment potential) and chemotypes (scaffold hopping potential). A framework for retrospective evaluation of virtual screening, which runs with different descriptors, was established on the

basis of the COBRA (Collection of Bioactive Reference Analogues) database [14], employing Euclidean distances for metric descriptors and the Tanimoto coefficient for fingerprints [15]. The COBRA database currently contains 12,642 manually curated bioactive compounds with 980 target protein subtype annotations. For 170 selected targets with a minimum of 20 annotated active ligands, every compound annotated as active was selected as a query and compared to all other compounds in the database in terms of similarity or distance of the respective molecular description, finally yielding sorted screening lists with the most similar or the least distant compounds sorted to the top.

The receiver operating characteristic (ROC)-related BEDROC score was selected for actives-retrieval benchmarking [16], while the Murcko scaffold [17] diversity among the set of actives within the top 1% of respective screening lists served as measure for scaffold hopping potential (Figure 7.3). For our study, the alpha level of the BEDROC method was set to 160.9, which corresponds to the top 1% of the screening list, contributing 80% of the score and emphasizing early enrichment.

For benchmarking purposes, we used a representative set of descriptors and fingerprints. First off, “Morgan” fingerprints closely related to extended-connectivity (radial) fingerprints are based on radial assessment of nonpredefined potentially infinite molecular fragments [18], whereas the “AtomPair” descriptor can be seen as a CATS predecessor merely counting the occurrence of all pairs of atoms in a

![Figure 7.3](image)

**Figure 7.3** Virtual screening benchmark facilitated by BEDROC score distributions and scaffold hopping potential for selected proteins targets on the COBRA database regarding various descriptors. Scaffold diversity in the set of retrieved actives among the top 1% screening entries is depicted inside the boxplots in the form of the number of unique scaffolds as well as a ratio thereof to the number of retrieved actives (in percent).
compound [4]. The “MACCS” keys represent substructure-based fingerprints, while the “RDKit” fingerprint implements a Daylight-like fingerprint based on hashed molecular subgraphs [19]. Latter fingerprints and descriptors were calculated using the open-source cheminformatics software package RDKit [20]. Finally, the “MOE2D” descriptor consists of a standardized vector of physicochemical properties provided by the Molecular Operating Environment (v2011, Chemical Computing Group, Montreal). At this point, we analyzed two versions of CATS vectors, CATS1 and CATS2, where CATS2 distinguishes lipophilic from aromatic atoms during typing, thereby resulting in more pharmacophore-type pairs and a higher dimensionality of the descriptor than CATS1, which lacks the aromatic atom type.

The considered descriptors were sorted according to decreasing performance in terms of the median of nonnormally distributed BEDROC scores from left to right in Figure 7.3. Aside from the comparison of CATS2 to the MOE2D descriptor, every decrease in performance due to a change of the descriptor is significant (Figure 7.4). According to the resulting dendrogram, the descriptors can be classified into four actives-retrieving performance categories: Morgan/AtomPair > CATS2/MOE2D > MACCS/CATS1 > RDKit fingerprint. Due to the fact that none of the descriptor metrics performed outstandingly, it might be advisable to select

Figure 7.4 Heatmap depicting p-values derived from statistically comparing BEDROC score distributions of different descriptors, utilizing the Wilcoxon rank sum test with an alternative hypothesis of row-denoted descriptors having greater distributions than column-denoted descriptors [22]. Eggshell coloring indicates significantly greater distributions (*/*/*/***), while orange and red colors indicate no significant difference ([0.05,0.99] and 1, respectively). Clustering the p-values with Ward’s method resulted in the depicted descriptor dendrogram [23].

2) MACCS-II, 1984. MDL Information Systems/Symyx, Santa Clara, CA.
one method from each group for similarity searching and compare ranked results lists, for example, by data fusion [21].

Scaffold hopping potential was determined by $s$, the median of differing scaffolds within the set of $n$-retrieved actives in the top 1% of the screening lists (Figure 7.3, numbers above the boxplot median lines) and $r$ (Figure 7.3, percentages below the median boxplot lines), which gives the ratio of $s$ to $n$ in percent. While $s$ correlates to the BEDROC scores with regard to the different descriptors, $r$ unveils the CATS descriptor in both implementations as the most potent descriptor for scaffold hopping among the molecular representations compared. We wish to point out that this outcome might differ in other benchmarking exercises and is critically influenced by the structural diversity of the underlying screening compound pool.

CATS1 and CATS2 descriptors were employed in the overall benchmarking with their (default) optimal parameters determined by variations of autocorrelation path length as well as types of scaling (Figure 7.5). For both descriptors, “types scaling,” which mitigates the potential dominance of individual pharmacophore types, performs superior to no scaling or scaling by compound size. Selecting a correlation path length of at least eight bonds for types-scaled descriptors delivers actives-retrieval performance not significantly differing from enrichment obtained with the optimum path length of 14 bonds.

Although state-of-the-art radial fingerprints outperformed CATS descriptors in terms of the number of actives retrieved, CATS ratify their intent of design by delivering the overall highest ratio of diverse scaffolds among retrieved actives. As the CATS2 extension significantly improves actives-retrieval performance with a

![Figure 7.5](image-url)  
*Figure 7.5* Median BEDROC performance of various parameterizations of CATS descriptors, resulting from virtual screening runs for selected targets on the COBRA database. The path length was varied from 1 to 17 as well as 43 bonds, while three different scaling types were applied, namely, pharmacophore types scaling, size scaling, and no scaling.
negligible loss of scaffold hopping potential, we recommend the usage of the extended version imperatively employing types scaling.

The outcome of our limited benchmark is in agreement with a large-scale systematic analysis of 2D fingerprint methods by Sherman and coworkers, who conclude “( . . . ) if the objective of a screen is to identify novel, diverse hits, then a less specific atom-typing scheme may be more appropriate” [24]. The CATS representation of molecular graphs and pharmacophoric features serves exactly this purpose of finding new chemotypes. When using the descriptor, one should not expect the highest possible enrichment of actives among the top-scoring virtual hits, but can anticipate surprising new ideas for synthesis and activity testing.

7.3 Prospective Scaffold-Hopping Applications

CATS descriptors have been successfully used in prospective similarity searching, aiming at finding scaffold hops from known drugs to innovative chemotypes, and combinatorial compound optimization. Here, we highlight some prototypic applications.

In a study performed at Roche, compound 1 emerged as a potent ($K_i = 2.4$ nM) A$_{2A}$ purinergic receptor antagonist from similarity searching in a virtual combinatorial library [25]. The underlying scaffold (Scheme 7.1, first panel) was completed by chemically suitable side chains to yield a pool of candidate compounds. Then, CATS descriptors were computed for all compounds and a self-organizing map (SOM) method was used to pick an optimal side chain decoration.

Compound 2 (clopimozid) represents the historically first scaffold hopping example with CATS [3]. Starting from a known drug as query (mibefradil, a cardiac T-type Ca$^{2+}$ channel blocker) clopimozid was identified as a nanomolar inhibitor of T-type Ca$^{2+}$ channels by similarity searching in a corporate compound database.

Researchers at Novo Nordisk used an implementation of CATS lacking lipophilic–lipophilic pair counts for similarity searching. In a biochemical high-throughput screening campaign, they identified an inhibitor of glycogen synthase kinase that they used as query for similarity searching in combinatorial libraries and successfully synthesized compound 3, a potent inhibitor of the target enzyme (IC$_{50} = 1.2$ μM) [26].

A series of inhibitors of 5-lipoxygenase, including natural-product-derived compound 4 (IC$_{50} = 0.8$ μM), were found by similarity searching [27]. Natural products and natural-product-derived combinatorial compound collections were virtually screened using 43 known 5-lipoxygenase inhibitors as queries. From the resulting list of candidates, 18 molecules were cherry-picked, ordered, and tested in a cellular and a cell-free assay system for inhibition of leukotriene synthesis. Two of them exhibited submicromolar inhibitory activity.

De novo design, that is, similarity searching in a virtual compound collection that is constructed by fragment assembly, is another domain of application for CATS descriptors. Two such examples are presented in Scheme 7.1. In each of these
studies, a known drug served as a template for computational compound
generation. The design algorithm used (TOPAS software [28]) assembles new
virtual molecules from a large stock of fragment-like molecular building blocks and
simple reaction schemes. During the compound construction process, each

Scheme 7.1 Examples of compounds that were
identified using CATS descriptors for (i) scaffold
decoration (1: \( A_{2A} \) purinergic receptor
antagonist, \( K_i = 2.4 \) nM); (ii) similarity
searching in compound databases (2: T-type
\( Ca^{2+} \) channel blocker, \( IC_{50} < 1 \) \( \mu \)M; 3: glycogen
synthase kinase inhibitor, \( IC_{50} = 1.2 \) \( \mu \)M; 4:
inhibitor of 5-lipoxygenase, \( IC_{50} = 0.8 \) \( \mu \)M; (iii)
template-based de novo design (5: \( Kv1.5 \)
potassium channel blocker, \( IC_{50} = 0.5 \) \( \mu \)M; 6:
cannabinoid 1 receptor inverse agonist,
\( K_i = 0.3 \) \( \mu \)M).
intermediate is compared for similarity to the template using CATS descriptors. Ultimately, the algorithm produces designs featuring new scaffolds but conserved pharmacophore patterns. Compound 5 is a de novo-designed Kv1.5 potassium channel blocker ($IC_{50} = 0.5 \mu M$) [29], and compound 6 an inverse agonist of cannabinoid 1 (CB1) receptor ($K_i = 0.3 \mu M$) that was further optimized to yield a clinical candidate [30,31].

7.4 Conclusions

Topological pharmacophore feature pair descriptors like CATS represent molecules in a way that enables scaffold hops. Numerous practical applications have confirmed the usefulness of this concept. While there may always be a context-sensitive, better-suited molecular descriptor for the purpose of actives enrichment, CATS might represent a method of choice whenever a balanced ratio of actives and underlying diverse scaffolds is desirable. Apparently, there is no universally applicable best molecular representation and similarity metric that are perfectly suited for the retrieval of both actives and diverse scaffolds at the same time. This is in part owed to the fact that compound databases and collections are often biased toward certain drug targets or chemotypes and contain series of close structural analogs. Keeping in mind that CATS is designed for the purpose of scaffold hopping, by visual inspection of hit lists one may expect to find surprising new compound structures with desired bioactivity.

References


8 Reduced Graphs
Kristian Birchall

8.1 Introduction

In essence graph reduction attempts to simplify objects, distilling their information and facilitating processing. The most familiar example of this concept is in encoding biological macromolecules such as proteins and sugars; highly complex structures can be simplified into a linear sequence or branched diagram, indicating the connectivity between the subunits. This chapter explores the concept of graph reduction as applied to small molecules. The term graph reduction comes from the branch of mathematics called graph theory, which deals with the description and analysis of objects known as graphs that consist of nodes connected by edges. This is of particular relevance, since fundamentally a chemical structure can be described as a chemical graph (CG), that is, a collection of atoms (nodes) connected by bonds (edges). Graph reduction is the procedure whereby substructures are condensed into individual nodes, while the connectivity between the substructures is retained, as illustrated in Figure 8.1. For chemists this is an intuitive and familiar notion underlying the principle of functional groups and systematic naming. For example, paracetamol (systematic name N-(4-hydroxyphenyl) acetamide) contains a hydroxyl group, a phenyl ring, and an acetamide group; each group can be represented as a node in a reduced graph (RG), as shown in Figure 8.1b. The name paracetamol itself is derived from a contraction of the constituent chemical groups (para-acetylamino phenol) containing an acetyl, an amino, and a phenol group, which can result in an alternative RG, as shown in Figure 8.1c. The partitioning of a CG into a RG is governed by a set of algorithmic rules, referred to as the reduction scheme. There are many possible reduction schemes, resulting in different RGs that may be smaller, larger, or even of the same structure. Furthermore, there are many possible variations in the level of information encoded by the node type such that the same RG can be described with differing levels of detail. The choice of scheme and level of graph reduction are key determinants of the descriptive power of RGs and will be explored in Section 8.2.

In the context of scaffold hopping, the primary attraction of graph reduction is that several molecules with different chemical structures can be recognized as
similar by virtue of having the same reduced graph (Figure 8.2). Just as there are a range of methods that can be used to compare molecules, there are a range of methods that can be applied to calculate a similarity value between RGs. While there is overlap in the methods used for CGs and RGs, there are important differences due to the smaller more information-dense nature of the RG. There will be a discussion of approaches for calculating the similarity between RGs in Section 8.3.

The emphasis on a functional rather than structural encoding means that the RG can be regarded as a topological pharmacophore. Not only does this facilitate scaffold hopping, it also makes RGs attractive for use in related applications such as clustering and analysis of structure–activity relationships, although these topics are beyond the scope of this chapter.

Figure 8.1 The unambiguous molecular structure of paracetamol can also be referred to as a chemical graph, where atoms correspond to nodes and bonds correspond to edges (a). The remaining depictions are reduced graphs derived from this structure. Depending on the reduction scheme, it is possible to generate different RGs with the same number of nodes (b and c) or with different numbers of nodes (d and e). Furthermore, the resulting RG can be described with differing levels of detail as encoded by the node types (f and g).
Reduced graphs are generated from chemical structures by application of algorithmic rules, collectively known as a reduction scheme, that identify which atoms are grouped into which nodes. A variety of reduction schemes are possible, the choice of which determines the size and structure of the resulting RGs, as illustrated in Figure 8.3. The ring/nonring reduction scheme simply involves cutting all bonds between ring and nonring atoms before assigning a node type to the resulting fragments. A variation on this involves decomposing fused ring systems to the smallest set of the smallest rings rather than having a single node representing an entire fused ring system. The carbon/heteroatom reduction scheme collapses contiguous carbon atoms into a node, with contiguous heteroatoms also being reduced to a single node. The homeomorphic reduction scheme assigns nodes to terminal atoms and branch atoms (degree > 3), resulting in a RG that emphasizes topology to a greater extent than the other exemplified reduction schemes, although each reduction scheme retains something of the topology of the chemical graph.

Figure 8.2  The many-to-one relationship between chemical graphs and reduced graphs can be very useful in recognizing structurally diverse compounds that share a similar pharmacophore.
8.2.2 Node Labeling

For a given graph reduction scheme, it is possible to add further detail to the description by implementing different node labeling schemes. For example, further detail could be added to the description of the ring/nonring reduction by assigning different node types (labels) according to whether a ring is aromatic or aliphatic. Other features such as H-bonding potential can also be encoded and combined in a hierarchical manner in order to generate RGs that are increasingly discriminative. Gillet et al. [1] applied this type of labeling scheme to RGs resulting from a ring/nonring reduction scheme, and an illustration of this is provided in Figure 8.4. It is worth emphasizing that the RG structure remains constant; only the node labels are changed. As the level of detail in the node description increases, so does the discriminative power of the RG. At the least detailed level of description (Level 1), structures (a)–(f) give rise to the same RG; at Level 2 only structures (a)–(e) give rise to the same RG; at Level 3 only structures (a)–(d) give rise to the same RG; at the most detailed level of description Level 4, only structures (a)–(c) give rise to the same RG. However, even at the highest level of discrimination, there is still a
considerable variation in the types of structures that would be considered identical
in terms of their RGs.

8.2.3 Sheffield Implementations

The R/F reduction scheme developed by Gillet et al. initially focussed just on
encoding rings and H-bonding features, since these were perceived to be key
groups in biological target recognition. The procedure for generating the RG can be
summarized as follows: identify H-bonding atoms using SMARTS, recursively
remove terminal non-H-bonding atoms, assign nodes for each ring in the smallest
set of the smallest rings, assign linker nodes for connected isolating carbons (C
atoms not doubly or triply bonded to any heteroatom), the remaining acyclic parts

![Figure 8.4](image-url)
form feature nodes that are then connected to the rings as in the original CG. This algorithmic procedure enables processing of unusual chemical structures that may otherwise be missed in the earlier dictionary-based approach taken by Takahashi et al. [2], where RG nodes are only encoded for predefined substructures. Even so, graph reduction can occasionally fail when processing extremely complicated CGs such as the alkaloid vinblastine.

Later work went on to explore variations on the graph reduction procedure developed at the University of Sheffield by Gillet et al. For example, Barker et al. [3] investigated the encoding of terminal non-H-bonding groups to reflect their importance as a hydrophobic pharmacophore point. Barker et al. also investigated the encoding of isolating carbon atoms as individual linker nodes rather than the more generic representation, where adjacent isolating carbon atoms are grouped into a single linker node. Harper et al. [4] also built on the work of Gillet et al. by recognizing positively and negatively ionizable features in addition to extending the ring/feature combinations to give 18 different node types: aliphatic ring, aromatic ring, or acyclic, each in combination with one of the six features assigned in the following order of precedence: positively ionizable, negatively ionizable, joint H-bond donor and acceptor, donor, acceptor, or non-H-bonding. More recent work by Birchall et al. [5] had a similar aim to that of Harper et al. in terms of the range of resultant pharmacophoric node types, although with fundamental algorithmic differences to facilitate a simple, rapid implementation in Pipeline Pilot.1)

Fragments are generated by recursively cutting all nonterminal acyclic single bonds with the following three exceptions to ensure chemically meaningful results: sp3 carbon to sp3 carbon bonds, heteroatom to heteroatom bonds, and heteroatom to sp2 carbon bonds. Each fragment is represented by a node that is reconnected according to the topology of the CG and assigned a type according to the fragment properties. An important difference compared to earlier RGs is that this results in a more generic representation, where fused ring systems are encoded by a single node. Options were added to provide greater control and flexibility over the reduction scheme by allowing fused rings to be encoded as multiple nodes and by using SMARTS definitions to control the fragmentation procedure.

8.2.4 Extended Reduced Graphs

The extended reduced graph (ErG) approach developed by Stiefl et al. [6] is similar to that of the R/F(4)/T approach used by Barker et al., whereby rings and H-bonding features are encoded in addition to terminal features. However, the ErG approach also encodes charged features and hydrophobic features explicitly. Furthermore, the method differs significantly in the way it deals with encoding rings and molecular topology. The procedure is illustrated in Figure 8.5 and can be summarized as follows: encode nonring-charged and H-bonding features, encode hydrophobic endcaps, create ring centroids, retain ring feature atoms, retain ring

fusion atoms, connect retained features to centroids, connect nonring features to centroids via the shortest atom path, and remove remaining ring atoms. Although this results in a more complex looking graph than those produced using the approach developed by Gillet et al., the beneficial consequences of this are that positional information is better conserved and interfeature distances in the original molecule tend to be more accurately represented. Furthermore, separating the encoding of ring features from the ring itself permits similarity to be reflected between rings of different feature types. For example, while the Gillet et al. approach would encode pyrrole and phenyl rings as different individual nodes, the ErG approach would result in an aromatic node joined to a donor node for pyrrole, thus presenting some commonality with the single aromatic node resulting from a phenyl group. Another potential improvement is that fused rings such as adamantane are collapsed to a single hydrophobic node to prevent overrepresentation in the RG.

Whatever, the choice of reduction and labeling schemes, inevitably there is a loss of information in going from the CG to the RG. The more generic representation of a RG compared to a CG allows a greater diversity of structures to be recognized as similar, which tends to be beneficial for scaffold hopping. However, as the representation becomes more generic, a greater number and diversity of structures will be recognized as similar and so it can become difficult for the user to identify the most interesting molecules from those that would be considered too diverse to be of interest. The balance between generality and relevance is a key issue that can be tuned by the choice of reduction and labeling scheme as well as by the approach used for calculating similarity between the RGs, which is the subject of the next section.

8.3 Comparison and Usage of Reduced Graphs

Following the seminal work of Gillet et al., RGs have been represented computationally using heavy metal atoms to encode different node types; for example, zinc was arbitrarily chosen to encode a linker node and scandium to
encode a featureless aromatic ring node. Connections between the RG nodes are by default single bonds, although Harper et al. later made use of the double bond to encode ring fusion (Figure 8.7). Despite the unusual chemical implications, this enables processing of the RGs by conventional chemical software, including structure depiction, pseudo-SMILES encoding, and fingerprint calculation.

8.3.1
Conventional Fingerprinting

Just as there are a variety of ways in which CGs can be compared, a number of different approaches have been taken in comparing RGs. The most commonly used approach for calculating similarity is based on fingerprints. Gillet et al. first tried using Daylight fingerprints (DFPs)\(^2\) calculated directly from the heavy atom pseudo-SMILES. Performance in simulated virtual screening experiments across a range of activity classes found that although the retrieval rates were slightly below those of the DFP applied to conventional CGs, the RGs tended to identify actives that were different to those found using the CGs. Significantly, many of the actives retrieved were of very low structural similarity to the queries, demonstrating the potential of RGs for scaffold hopping. An example of the structural diversity of retrieved actives is shown in Figure 8.6. The study investigated the effect of using different reduction and node labeling schemes and found that performance varied between activity classes such that no single approach was consistently best.

![Figure 8.6](http://www.daylight.com)

While DFPs calculated from RGs provide a simple means of using RGs in similarity searches, there are a number of reasons why they are not ideal. DFPs were designed for use with CGs, which are typically much larger than RGs. A consequence of this is that when applied to RGs the fingerprints are quite sparse, which can distort the similarities that are calculated between them. Furthermore, the smaller size of RGs means that identical paths may occur multiple times and this is not accounted for in DFPs that only encode the absence or presence of a fragment. A final limitation of using DFPs is that they are hashed and so cannot be easily interpreted, although this is not specific to their usage in the context of RGs.

### 8.3.2 RG-Specific Fingerprints

Barker et al. investigated a number of different possible fingerprinting schemes that were applied to the RGs developed by Gillet et al. The descriptors were based on the atom-pair approach published by Carhart et al. [7], taking the general form of node–distance–node. Two types of distances were investigated: the number of edges between the nodes in the RG and the number of bonds that separate the node fragments in the CG. Node path descriptors that encode specific sequences of nodes (up to a maximum of three to keep within the fingerprint size limits) were also investigated. In each case, each descriptor sets a different bit in the fingerprint, ensuring its interpretability, something afforded by the smaller size of the RG compared to the CG. For the three different descriptor approaches, both binary (presence/absence) and hologram (frequency) versions of the fingerprint were investigated. In simulated virtual screening experiments using 11 activity classes extracted from the MDDR data set [8], it was found that none of the fingerprints was consistently the best performing over all activity classes investigated. However, there was only one of these classes in which the RG DFP performed best, showing as expected that RGs benefit from having their own specific fingerprinting scheme. The interpretable nature of the resulting fingerprints was exploited by Barker et al. in order to derive SAR information using recursive partitioning. This established technique involves statistical determination of the features that best separate a training set of actives and inactives. The fingerprint-encoded features identified two simple node–bond pairs that were able to classify accurately the actives and the inactives, but more interestingly were found to correspond closely to the established pharmacophore for the target in question; hence, illustrating the value of the pharmacophore-like description provided by the RGs.

Harper et al. also developed a node–edge pair fingerprinting scheme similar to that described by Barker et al., but with several modifications designed to improve performance. For example, feature frequency information was encoded by setting separate bits to indicate the number of occurrences of each feature. To

---

help encode molecular topology bits are set to define branch points in the RG and ring fusion in the CG. Other bits are set to encode the frequency of heteroatoms in the nodes, which helps to distinguish between fragments such as pyridine and pyrazine. A general limitation of fingerprint-based approaches is that the same features at slightly different distances will set no bits in common. To deal with this, bits are set to encode paths that are of length one less such that a node–edge pair with a path of length five will also set a bit for the equivalent node–edge pair of path length four. However, there is a second limitation that is not dealt with in the fingerprint. There is no concept of functional similarity between RG node types, that is, a joint donor/acceptor node will set completely different bits to a donor node, for example, even though they may be able to fulfill similar functions biologically and hence should be considered similar to some extent.

8.3.3 Augmenting Fingerprints with Edit Distance

To take better account of these limitations, Harper et al. developed an edit distance-based approach to assessing RG similarity. Edit distance is widely used to assess the similarity of text strings, from spelling correction in word processing to the assessment of gene or protein sequence similarity in bioinformatics [9]. The method measures similarity based on a weighted sum of costs associated with the operations that are required to interconvert two objects. These operations include insertions/deletions, which enable the approach to cope naturally with insertions and deletions when comparing RGs. Furthermore, assigning differing costs to the substitutions between different node types also allows the similarity between the different node types to be dealt with naturally. As detailed by Harper et al., there are some limitations of the edit distance similarity approach arising from the fact that it is based on maximal paths rather than graph comparisons. To a large extent these are overcome by using a combination of the RG fingerprint and edit distance similarities. Similarly, the strengths of the edit distance approach help to compensate for the weaknesses inherent in fingerprinting approaches. An illustration of the sorts of false-positives arising from these limitations is provided in Figure 8.7.

Harper et al. evaluated the scaffold-hopping ability of the RG fingerprint/edit distance combination to see whether 10 CB1 antagonists could be identified as ranked in the top 100 of a data set containing approximately two million commercially available compounds. A tough but more realistic test compared to evaluations based on the MDDR data set, where several hundred actives are seeded into less than 200 000 compounds, with assessment based on retrieval in the top 1%, that is, over 1000 molecules. Both Daylight fingerprints and the RG approach each identified three actives and there was no overlap between these. This is in agreement with previous studies that found that the strength of RGs is in their ability to identify complementary sets of actives to those found using standard similarity searching.
8.3.4 Extended Reduced Graph Fingerprints

The ErG approach also uses fingerprinting as a descriptor for use in similarity searching. Fingerprint generation resembles that used by Harper et al., although while Harper et al. used a binary fingerprint and set bits to encode RG path frequency, Stiefl et al. used a hologram approach, where each bit encodes a count of the fragment frequency. A further difference is in the way they attempt to reflect the similarity between paths that differ by the insertion/deletion of a single atom. Harper et al. achieved this by setting bits to encode paths of length one shorter, while Stiefl et al. increments the bits for paths of length one shorter and one longer by a small amount (e.g., 0.3). Increasing or decreasing this amount allows the user to retrieve more or less distantly related compounds, respectively. In simulated

Figure 8.7 Examples highlighted by Harper et al. to show the sorts of false-positives that can arise from the limitations of using edit distance and RG fingerprint approaches in isolation. The RGs are depicted using their heavy metal pseudo-SMILES, where Sc encodes featureless aromatic, V encodes aromatic H-bond acceptor, Ni encodes acyclic H-bond acceptor, Y encodes aliphatic positively ionizable, and double bond encodes ring fusion. The path-based nature of edit distance approach results in an artefactual similarity of 1 for the pair of reduced graphs shown in part (b). Despite topological disparities, the query molecule ranked the hit 44th on the basis of the RG fingerprint alone. The hit was ranked a more reasonable 1407th, when using the edit distance measure in combination with the RG fingerprint.
virtual screening experiments on 11 activity classes extracted from the MDDR data set, the retrieval rates obtained with ErG were found to be as good as or better than DFPs in all but one activity class. However, in each activity class ErG was able to retrieve actives from a greater number of structural classes; assessed based on preclustering using DFPs. While both approaches were able to uniquely retrieve some actives, a greater number of unique hits were identified when using ErG. To illustrate the scaffold hopping potential of the ErG approach, Figure 8.8 shows representative examples of the structural diversity of 5HT reuptake inhibitors retrieved in the top 1% of the database, when searching with the depicted query. In this case both ErG and DFP approaches identified 16 actives in the top 1% of the database, with only a single active in common between the two hit lists. In addition to the high degree of complementarity, it is notable that the ErG approach identified a large range of structurally diverse actives, whereas the actives retrieved using DFPs tended to be analogs of each other, thus limiting hit list diversity.

While the details of exactly what makes one method perform better than another in different activity classes is a complex multifactorial issue, the authors highlighted that performance of the RGs tended to be correlated with feature density, that is, a higher number of RG features relative to the size of the molecule were observed in cases where the RGs outperformed CGs and vice versa. This reflects an important general point: RGs will tend to excel in activity classes where there is a sizeable well-conserved pharmacophore that accounts for a significant proportion of the molecular structure, regardless of the structural diversity of the pharmacophoric points. Conversely, smaller feature-sparse RGs are more sensitive to

Figure 8.8  Representative examples of the structural diversity of 5HT reuptake inhibitors retrieved using the query structure. Several striking examples of scaffold hopping are achieved using ErG, while far fewer scaffold hops were achieved when using DFPs.
changes in query structure and so matching actives tend to be obscured by the
greater number of chance matches against database molecules. Distilling
pharmacophoric information by graph reduction will inevitably also enrich features
that may not directly contribute to binding; for example, in peptidic ligands the
frequent occurrence of the backbone amide is the predominant pharmacophoric
feature, even though it may simply be acting as a linking rather than interacting
group. Another example of wider relevance is where groups such as morpholine or
hydroxyl may be appended at a variety of solvent-exposed positions simply to
improve solubility. Although it is true that such information is a diluting factor
both when dealing with RGs and CGs, the effect is exacerbated in RGs since the
number of features is much lower and so the relative dilution is much greater.

Stiefl and Zaliani [10] developed an interesting extension to the ErG work that
emphasizes the flexibility of the approach and the importance of incorporating
knowledge to improve performance. Based on knowledge of the drug–target
interaction gained from experimental data, all molecular features that could
potentially form similar interactions with the target are treated with greater
significance. This is implemented by weighting (increasing the value of) the ErG
fingerprint in the bit positions that encode those features. The performance in
simulated virtual screening experiments when taking this approach improved upon
that of ErG and DFPs. However, it should be noted that this approach is dependent
on the availability of experimental data and on the correct interpretation of the
types of molecular features that mediate the drug–target interaction.

8.3.5
Graph Matching Approaches

While most work on RGs have used fingerprints for making comparisons, graph
matching methods have also been investigated, initially by Takahashi et al. and later
on a more thorough test set by Barker et al. [11]. Graph matching is a
computationally intensive procedure, where the time taken can be exponentially
proportional to the size and complexity of the graphs being matched. Consequently,
the smaller size of RGs compared to CGs allows significant improvements in
speed. A notable attraction of the graph-based approach is that it tends to give a
more intuitive reflection of similarity. Topological connectivity relationships are
maintained unlike in fingerprints where node pair information is abstracted
with some loss of contextual information that can lead to false-positives (e.g.,
Figure 8.7).

Barker et al. represented the reduced graph as a fully connected graph in which
the edges represent bond distances in the original chemical graph. They
investigated the use of maximum common subgraph techniques to calculate the
similarity between pairs of reduced graphs. In simulated virtual screening
experiments on 11 activity classes extracted from the MDDR data set, the authors
showed that this method is capable of retrieving greater numbers of actives than
either DFPs or RG fingerprints. As previously found by Gillet et al., the actives
retrieved using RGs were complementary to those retrieved by DFPs. Also in line
with previous observations, relative performance of the different methods varied according to the activity class, although overall the graph matching approach was best, exceeding that of the RG fingerprints previously described by Barker et al. The authors also demonstrated the scaffold-hopping ability of the graph matching approach based on their retrieval of an increased number of different Murcko frameworks [12].

### 8.3.6 Bioisostere Encoding

Birchall et al. also took a graph matching approach to investigate whether bioisostere information could be taken into account during similarity searching. Bioisosteres extracted from the BIOSTER database\(^4\) were often found to be encoded by the same node type, supporting the suitability of the RG encoding. However, there were also many cases where the bioisosteric fragments were not encoded as the same node type or even by the same number of nodes. The graph reduction and matching schemes were then modified to recognize and permit matches between instances of bioisosteric fragments, regardless of RG node type, thus enhancing the similarity between molecules containing such fragments. Simulated virtual screening experiments on six activity classes extracted from the WOMBAT database\(^5\) found that although this approach resulted in modest improvements in scaffold hopping potential, assessed based on the number of different Murcko frameworks retrieved, there is a significant trade-off in terms of the number of inactives that are retrieved. Further investigation found that the main limitation appears to be in complexities of bioisosteric equivalence, the reliability of which depends on both the intramolecular environment and the target interaction environment. It is important to note that this limitation is not specific to the RG approach and would affect any other approach permitting matching of such equivalences based on current data. However, the difficulty of identifying reasonable limits of bioisosteric equivalence parallels one of the two key challenges with using graph matching applied to RGs. Namely, that the binary nature of node type matching is simply too coarse to truly reflect the spectrum of similarities that result when RG nodes are compared. Pharmacophoric node type classification is a catchall encompassing a range of changes in size, shape, flexibility, context, connectivity, magnitude of polarity, and relative orientation of polarity, as elaborated in Figure 8.9. Thus a node similarity based on a comparison of multiple such properties would seem a more appropriate way of deciding whether a pair of overlaid nodes should match, or indeed to what extent (similarity) they match. The other key challenge is centered on the partitioning of the RG, which is a crucial determinant of how the nodes are aligned. A single fixed partitioning is likely too rigid to work well in all scenarios, as illustrated in Figure 8.9, and so a dynamically adaptive approach may be more suitable.

---


Two key challenges when using RGs to assess molecular similarity. (a) The node partitioning issue. (b) The node equivalence issue. Using a given reduction scheme in part (a), the molecules are partitioned into nodes as indicated by the shading. The resulting poor alignment could be rectified by merging a2 and a3 to a single node and merging a4 and a5 to a single node, hence allowing a better reflection of similarity between molecules A and B during graph matching. The node equivalence issue is illustrated by considering the diversity among the terminal fragments in part (b) (where R is the attachment point), which would result in a single RG node broadly of the type hydrophobe, since they are all devoid of H-bonding or charged features. Even just considering the acyclic fragments along the top row, there is clearly a large variety in terms of size, shape, and flexibility, which would vary their ability to make hydrophobic interactions in a given pocket; for example, while the difluoromethyl group and t-butyl group are both essentially globular hydrophobes, the t-butyl group occupies a much larger volume. Furthermore, the t-butyl group is electron donating while the difluoromethane is electron withdrawing, which can produce significant differences in the electronics of a neighboring group. Even where there are similarities in one property, there are differences in another; for example, the propyl alkane and alkyne are both unbranched groups composed of three carbon atoms, but they differ significantly in terms of flexibility and electronics. Similarities can sometimes be greater between classes than within classes; for example, the cyclopropyl group more closely resembles the t-butyl group in terms of shape and volume than does the cyclohexyl group. There are also differences between the dimethylthiophene and dichlorophenyl fragments, depending on context; while they are both flat hydrophobes, the dimethylthiophene is much more electron rich than the dichlorophenyl fragment and so would be better-suited to interact with cations in a biological target. Furthermore, the ortho-substitution of the chlorophenyl would result in significant conformational twist in the context, where the fragment is directly attached to another bulky group such as an aromatic ring; the dimethylthiophene is much less susceptible to this effect by virtue of the different substitution pattern. The examples provided are the simplest scenario, since when charged or H-bonding features are involved, the relative location, directionality, and magnitude of the polarity also play key roles in determining equivalences, particularly since such features are often involved in mediating specific interactions with a target protein. Furthermore, the question of reasonable equivalence becomes increasingly complex when the node is acting as a linker or a branch point, since relative positioning of the attachment points and flexibility of the node fragment can produce drastic differences in the positioning of features, as could easily be imagined if a second or third R group was appended to the exemplified fragments.
8.4 Summary

Graph reduction provides a simple means to distil pharmacophoric information for use in high throughput virtual screening. By altering the rules used for graph partitioning, node type assignment and node type matching, reduced graphs provide the flexibility to allow the recognition of increasingly structurally distinct equivalences. However, this must be balanced against the degree of information loss inherent in graph reduction that may lead to the recognition of unreasonable equivalences. The key is in deciding what constitutes a reasonable equivalence and when. Understandably there is no single approach that will give the best retrieval rate for every activity class, however, several studies have proven the scaffold-hopping ability of RGs and the fact that they tend to retrieve actives that differ from those identified by conventional chemical structure fingerprints. RGs also have uses outside scaffold hopping, mainly in clustering and SAR analysis where their simplified pharmacophoric nature, particularly in combination with intuitive graph matching approaches, has been of great value. The interested reader is directed to the following references that are beyond the scope of this chapter [13–15].

References


9
Feature Trees

Nathan Brown

9.1
Introduction

Many scaffold hopping methods function through the introduction of a controlled fuzziness into the descriptors used [1]. Chapter 8 introduced the concept of reduced graphs, whereby molecular graphs are abstracted further by subsuming groups of atoms into nodes with connectivity being preserved between these nodes with representative edges [2]. An extension to the reduced graph was introduced in 1998 and was called Feature Trees and was specifically designed for evaluating molecular similarity in virtual screening applications [3].

Due to the introduction of fuzziness to the description, Feature Trees are very appropriate for scaffold hopping and also for the more general bioisosteric replacement problem. In this chapter, the method by which Feature Trees are generated will be presented and the principles of comparing Feature Trees will also be discussed briefly. Finally, examples of the application of Feature Trees in virtual screening both extant and virtual chemistry spaces will be summarized based on published work from the literature.

9.2
Feature Tree Generation

The Feature Tree for an individual molecule is generated using a number of abstraction rules to determine which atoms or groups of atoms will be subsumed into a node. It is important to first identify the ring systems since it is necessary for each of these to be collapsed into nodes to ensure an appropriate tree is generated. In this first step, the identified rings are subsumed into nodes and the subsequent graph called the cycle graph. The acyclic linkers and substituents are then abstracted as Feature Tree nodes. All nonterminal, acyclic atoms and their adjacent pendant atoms are also collapsed into Feature Tree nodes, giving the final tree (Figure 9.1).

Each node in the generated Feature Tree is assigned properties relative to the properties of the their constituent building blocks. Two different property types are
assigned: steric and chemical features. Two steric feature properties were defined in the original definition: the number of atoms in the Feature Tree node and an approximated van der Waals volume. The chemical features are based on the FLEXX interaction profile [4] and encode a number of pharmacophoric features: H-bond acceptors and donors, aromaticity, and hydrophobicity.

9.3 Feature Tree Comparison

Three Feature Tree comparison algorithms have been developed since the first publication in 1998: Split-Search [3], Match-Search [3], and Dynamic Match-Search [5]. The reader is referred to these articles for the precise definition of the algorithms. Each algorithm returns a set of optimal matches using a weighted-sum of the steric similarity and the chemical similarity. The resulting similarity for a match, \( \text{sim}(m) \), is referred to as the direct similarity of the subtrees. The similarity of the two parent Feature Trees is then calculated as the weight average of all the similarities, weighted by the sum of the sizes of the subtrees and scaled by \( \mu \) times the total size of unmatched subtrees.
Retrospective Validation

In the original paper, Feature Trees was retrospectively validated using two data sets: a data set compiled by Briem and Kuntz [6] from the MDDR database and a further data set compiled by Lemmen and Lengauer [7] from the Protein Data Bank [8]. The data sets are summarized in Tables 9.1 and 9.2, respectively. Two structures from the MDDR set contained macrocycles, which Feature Trees cannot handle and thus were removed from the data set.

The Feature Tree representations were identified as effective at being able to retrieve hits with low structural similarity to the query molecules in a number of cases. For ACE, HMG, and PAF, Feature Trees were shown to retrieve roughly the same numbers of actives compared with a similarity search using Daylight fingerprints [9], but in each case Feature Trees and Daylight fingerprints identified different hits on the order of 50% of the hit lists. For TXA2 and 5HT3, Feature Trees outperformed Daylight fingerprints substantially. Feature Trees identified

Table 9.1 MDDR ligand data set.

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of ligands</th>
<th>Average number of atoms</th>
<th>Receptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>40</td>
<td>63</td>
<td>Angiotensin-converting enzyme</td>
</tr>
<tr>
<td>TXA2</td>
<td>49</td>
<td>56</td>
<td>Thromboxane A2 receptor</td>
</tr>
<tr>
<td>5HT3</td>
<td>52</td>
<td>48</td>
<td>5-HT3 receptor</td>
</tr>
<tr>
<td>HMG</td>
<td>114</td>
<td>64</td>
<td>HMG CoA reductase</td>
</tr>
<tr>
<td>PAF</td>
<td>136</td>
<td>70</td>
<td>PAF receptor</td>
</tr>
<tr>
<td>RND</td>
<td>579</td>
<td>57</td>
<td>Random selection</td>
</tr>
</tbody>
</table>

Table 9.2 PDB ligand data set.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Number of ligands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carboxypeptidase A</td>
<td>5</td>
</tr>
<tr>
<td>Dihydrofolate reductase</td>
<td>2</td>
</tr>
<tr>
<td>Endothiapepsin</td>
<td>5</td>
</tr>
<tr>
<td>g-Phosphorylase</td>
<td>4</td>
</tr>
<tr>
<td>Immunoglobin</td>
<td>6</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>8</td>
</tr>
<tr>
<td>Streptavidin</td>
<td>5</td>
</tr>
<tr>
<td>Thermolysin</td>
<td>12</td>
</tr>
<tr>
<td>Thrombin</td>
<td>4</td>
</tr>
<tr>
<td>Trypsin</td>
<td>7</td>
</tr>
</tbody>
</table>

1) Accelrys, Inc., 5005 Wateridge Vista Drive, San Diego, CA 92121, USA. MACCS Drug Data Report (MDDR).
more diverse hit matter with substantially different scaffolds, demonstrating their effectiveness at scaffold hopping.

9.5
Implementations and Applications

FTrees is available with a number of interfaces. An interface is available from BioSolveIT itself, called FTreesXL, and provides a standalone interface for running FTrees queries. In addition, FTrees is available from two third-party software vendors FTrees in PipelinePilot and FTrees in MOE. Finally, a web interface has been developed that allows the user to search the ZINC [10] database free-of-charge. All these packages are available from BioSolveIT.2)

Since the inception of Feature Trees, a number of applications have been discussed from pharmaceutical companies to perform searches using multiple queries (SanofiAventis) and to explore vast fragment spaces (Pfizer and Boehringer Ingelheim). The approaches that have been reported are discussed in the following sections based on reports in the literature.

9.5.1
MTree: Combinations of Query Molecules

There are dual challenges in drug discovery of not having sufficient experimental data in the early stages to develop quantitative structure–activity relationship (QSAR) models and also the degree of manual intervention in deriving pharmacophore models from limited ligand data. Therefore, Hessler et al. considered combinations of Feature Tree representations of molecules of interest to generate unified query molecules [5]. These FTrees are then unified into a single MTree to search in the Feature Tree space. The authors highlight that this is essentially a ligand-based topological pharmacophore representation (introduced with CATS vectors in Chapter 7). In this way, the methods capitalize on the lack of resolution of the structural description and are highly suitable to scaffold hopping.

9.5.2
Similarity Searching in Large Combinatorial Chemistry Spaces

Since Feature Trees are tree-based summaries of molecules with each node in the tree representing at least one atom, but typically functional groups, this makes them amenable to fragmentation and the subsequent combinatorial exploration of their chemistry spaces. The FTrees–FS system is an efficient implementation of the exploration of combinatorial chemistry spaces using dynamic programming [11,12]. The fragments are generated using the RECAP fragmentation scheme [13]

from the World Drug Index, rendering a set of 16,780 unique fragments and 21,386 linkers from known drugs. The set of fragments allows the combinatorial expansion of $10^{18}$ compounds.

The FTrees–FS system uses a variant of the Feature Trees representation and comparison algorithms to optimize potential analogs of a query compound. However, the analogs, due to the introduction of controlled fuzziness in the Feature Tree representation, will tend to be more structurally distinct but functionally similar to the query. In this way, the system is able to provide alternative yet plausible solution structures that may be considered by visual inspection, investigation of docking poses, or any other molecular modeling technique.

In 2008, Pfizer applied the system to perform similarity searches on 358 combinatorial libraries using the CoLibri (Compound Library Toolkit) system that permits access to more than $10^{12}$ possible final molecules [14]. The system was reported with the specific aim of conducting scaffold hopping campaigns \textit{in silico} using the Feature Trees representations of the products. This can be conducted with the Feature Trees software to systematically explore the virtual space, but without enumerating the entirety of the product space.

The system has subsequently been applied using the Boehringer Ingelheim scaffold and approved reagent list (1600 scaffolds and 30,000 reagents, respectively) to generate a potential library of $5 \times 10^{11}$ unique molecules, called BI CLAIM (Boehringer Ingelheim Comprehensive Library of Accessible Innovative Molecules) [15]. The system used the CoLibri software developed by BioSolveIT to encode the reaction schemes. The subsequent combinatorial fragment space can be searched, as before, using the Feature Trees system.

9.6 Conclusions

Feature Trees have been demonstrated to be of great value in virtual screening to identify potential scaffold hops. This has been achieved through the introduction of controlled fuzziness into the descriptor by making atoms and nodes generic features as part of the tree generation. The tree structure and novel matching algorithms have also permitted the Feature Trees to be compared rapidly, even being able to compare massive combinatorial fragment spaces.

Acknowledgment

N.B. is funded by Cancer Research UK Grant No. C309/A8274.

3) The World Drug Index is available by subscription from Thomson Reuters at thomsonreuters.com/world-drug-index/.
References


10
Feature Point Pharmacophores (FEPOPS)

Jeremy L. Jenkins

10.1 Similarity Searching in Drug Discovery

Drug discovery is fueled by two fundamentally different paradigms to identify lead compounds within a large chemical library. In hypothesis-free approaches such as high throughput screening (HTS), new chemical matter is identified based on the strength of activity in a relevant assay readout based on a significant statistical deviation from the sample distribution. In contrast, hypothesis-based lead discovery is driven by compound sets tailored by knowledge of a target or disease biology. For example, an X-ray crystal structure of a target protein may be used to dock an entire chemical library in order to provide a ranked list enriched with likely hits. In other cases, chemical structures that modulate a target are available from prior research on a target protein or a cellular phenotype. These are often referred to as reference compounds, probes, ligands, or tool compounds. Tool compounds may originate from literature assays, small-scale pilot screens, X-ray structures, endogenous ligands, or even from compounds that modulate homologous or orthologous proteins.

In the pharmaceutical industry, ligand-based hypothesis-driven drug discovery often takes the form of similarity searching. For example, the chemical structure of a tool compound or set of tool compounds may be used to rank order a pharmaceutical library using a chemical descriptor to computationally define the structures, combined with a metric to measure the similarity of a query compound to database compounds. Similarity searching is a frequently used, robust, and fast method to create focused sets of typically 10s to 1000s of compounds to test in well-based assays (96 384, or 1536 wells). Such sets represent a small fraction of pharmaceutical collections that typically contain more than one million compounds. This means that only one in 1000–10 000 compounds would be “cherry picked” from the collection and tested. Accordingly, to find new active molecules, the enrichment rate over random must be quite significant. In the author’s
experience across hundreds of HTS campaigns, one can expect at best one compound in 1000–10 000 to be bioactive in a random screen, depending on the nature of the target and the content of the screening library. Thus, if a typical focused screen of 1000–10 000 compounds is to be useful, we need at least a 10-fold enrichment over random to get sufficient hits for follow-up activities. Many chemical similarity methods have claimed this level of success, especially if applied in retrospective studies to preexisting data sets.

In practice, the utility of a similarity approach to a drug discovery program depends on whether the goal of the search is to provide structure–activity relationships (SAR) or to provide new chemotypes for the project. Traditional 2D similarity approaches are, by design, useful for generating SAR. In other words, they retrieve from databases compounds that resemble the query molecule and that can facilitate the understanding of how chemical features do or do not correlate with bioactivity. It is tautological but worth stating that a similarity method that enriches for molecules similar to a query molecule must simultaneously “de-enrich” for compounds dissimilar to the query molecule. Therefore, if the goal of a similarity search is to find new chemotypes or scaffolds to progress a project, 2D similarity methods may not be sufficient by themselves. Whether the similar molecules recalled by a similarity method actually constitute the same scaffold or a scaffold different from the query is a more complex question that requires a meaningful definition of scaffold (see below, Section 10.5). The goal of finding new chemotypes from a starting tool compound is known as scaffold hopping or chemotype hopping.

Many methods have been developed in the past decade to encourage the jump from one chemical framework to another without sacrificing bioactivity. These are the subjects of the chapters of this book. Scaffold hopping is an important topic for many reasons in the pharmaceutical industry. Among the many reasons, three stand out with some frequency. (i) Patents are typically written around Markush structures, an abstract form of a core scaffold with variable R groups or side chains that decorate the core. Markush structures help create intellectual property because they are often meaningful from a synthetic chemistry standpoint and patents can be designed to block competitor companies from making similar compounds. Similarity methods that jump away from a core framework will often “bust” such competitor patents. (ii) Project team compounds occasionally get stuck in “flat” SAR where changes to R groups do not dynamically alter bioactivity. In these cases, a new chemotype or series is desired to obtain more dynamic SAR. (iii) Some scaffolds may contain aspects that are toxic or are liabilities in clinical candidates or drugs. For example, specific combinations of hydrophobic features and a basic charged feature are often found in compounds that block the KCNH2 h-ERG channel in the heart. A method to hop to other bioactive molecules that do not contain these detrimental features is very important to the lead optimization in the process.

Simple graph-like representations of compounds or even physicochemical properties are extremely useful and robust tools for similarity searching. Some,
such as circular fingerprints, are quite capable in their own rite of scaffold hopping [1,2]. However, 2D descriptors can only indirectly represent the important features required for a compound binding to a protein. The compound–protein binding is driven by a mixture of forces that include enthalpic and entropic components. On the protein side, binding is influenced by protein pocket accessibility, binding site residue properties, pocket shape and flexibility, and solvation state. On the compound side, similar concepts apply: conformational shape and flexibility, atomic properties (e.g., charge, hydrophobicity), and desolvation requirements drive the likelihood that a compound will bind to a protein target. How does this relate to chemical similarity searching? In reality, chemical similarity searching is being employed by cheminformaticians as a proxy for “biosimilarity” searching; a similarity search is expected to highly rank compounds that not only look like the query molecule but behave like the query molecule, following the assumption that similar chemical structures tend to have similar activities. But to find compounds that behave like one another, we must remember that compounds are not 2D, not 3D, but exist in a flexible array of conformations in solution, spinning, rotating, and translating as they bounce off from water molecules, buffer, salts, solvents, and protein targets in an assay well. Compounds are 4D. They exist in both temporal and conformational space – a large leap from a 2D graph representation (despite the proven utility of the chemical graph). This is the complex environment in which we ultimately must assume that two compounds are similar in order to construct an orthogonal approach to 2D similarity.

10.2 FEPOPS: An Analogy to Image Compression

It is possible to computationally compute and store a representative array of conformational positions for a compound in solution. The number of potential conformers depends on the number of rotatable bonds and other steric and charge considerations. A large, flexible compound can create a combinatorial explosion of potential conformers. Worse yet, with respect to similarity searching, comparison of query compound conformers to the many billions of conformers that can be precomputed for an entire corporate collection becomes an intractable problem of scale. FEPOPS (feature point pharmacophores) is a 4D chemical similarity method inspired by this problem.

FEPOPS originated from concepts introduced at the Oxford University by Richards and coworkers, who attempted high-throughput computational docking to an entire protein surface [3]. Docking to the entire surface of a protein is also a high-dimensional problem that can be addressed by scale reduction. By using multiscaled representations of compounds, the authors could systematically eliminate the unlikely docking positions or “poses” to prioritize the more likely binding pockets. k-means clustering on the atomic coordinates of query molecules was carried out, starting with $k = 1$ to define the centroid, the geometric mean of all
atom coordinates. The centroid was docked to the surface of the protein using a grid-based energy method to find cavities for further docking. In the second iteration, $k$-means clustering of compound atoms of potential conformers was computed with $k = 2$, and the distance between the two points was recorded. The two points and the edge between them constitute a marginally larger representation to be docked on the pockets identified from the first iteration. The number of potential pockets and poses is thus reduced in the second round, and the process repeats, increasing the value of $k$ until docking solutions are exhausted or $k = n$, the number of atoms in the compound. The multiscale approach enabled accurate docking of test compounds to known protein targets in seven test cases [3]. The authors concluded that compounds could be docked using a smaller number of feature points than expected – typically four. In addition, the number of conformers necessary to represent the compound was actually quite small when using the multiscale concept. It appeared that typical flexible compounds in screening collections could be represented as a small set of rigid feature points. Based on these observations, the idea followed that similar scaled-down representations could be used to compare ligands, as opposed to ligand docking.

Clustering by $k$-means has been used in the computer graphics industry to cluster pixel colors in an image for compression. Image compression reduces the file size if, for example, an image needs to be electronically transferred. Components of an image are made up of red, green, and blue (RGB) colors stored as 3D vectors of floating numbers. Nearby vectors are similar if nearby colors are similar. The image size can thus be reduced by $k$-means clustering the vectors to reduce the complexity of the image to a smaller number of vectors. According to Allesandro Guisti, this “is part of a broader class named vector quantization, which has a number of applications in lossy data compression; indeed, reducing the number of colors in an image is useful to compress it: assuming to represent each color component with 8 bits, a small $100 \times 100$ image would need $10\,000 \times 24$ bits (29.3 kB); if we manage to acceptably represent the image using only 32 different colors, we can encode each with just 5 bits, and reduce the needed memory to 6.1 kB, plus the space required to define the 32 used colors ($32 \times 24$ bits).”

By analogy, if we think of a compound in four dimensions as the very large array of fully flexible conformers in solution, we can first represent the conformers as an overlapping set in a flattened image. For example, all possible adenosine triphosphate (ATP) conformers create a dense image of overlapping atoms and bonds (Figure 10.1). Then, as with image compression, the same atoms and bonds can be clustered to four $k$-means points per conformer (Figure 10.1, at right). This is a much simpler representation that covers much of the chemical conformational space that is largely redundant in the image of all overlapping conformers. This is how the multiscale concept can be used to create a scaled-down representation of compounds in 4D, and in turn, be used to rapidly similarity search a large chemical database.

1) http://www.leet.it/home/lale/joomla/content/view/12/26/.
Our rationale for creating a new chemical descriptor was to enable a fast, automated 3D similarity method that implicitly encoded ligand flexibility while alleviating the need to understand the important features for binding, the pharmacophore. Oftentimes, the bioactive conformation of a tool compound is unknown. Our method attempts to create fuzzy representations that reduce conformer space to a smaller scale to enable a small number of similarity measurements between any two molecules and to obviate the need for an a priori pharmacophore. The method would not replace rapid and reliable 2D similarity methods but serve as an orthogonal approach to lead finding and compound subset design.
The FEPOPS descriptor contains four points determined by \( k \)-means clustering, the atomic coordinates of a molecule [4]. Each point is encoded with properties of the atoms in its cluster. For example, in the first of the four \( k \) clusters, cluster 1, the atoms’ partial charges and atomic logP values are summed. A binary flag is set to 1 or 0 for presence or absence of hydrogen bond donors and acceptors. In a subsequent version of FEPOPS, a flag for positive or negative charges is added to the descriptor. Thus, the atoms of the molecule can be discarded and only the four points are kept with their summary properties. The four feature points are then sorted in ascending order of the sum of partial charges. For example, the most negatively charged of the four feature points is called feature point one, whereas the most positively charged feature point is called feature point four. This allows a simple canonicalization of the fuzzy structure to enable fast comparison between molecules. The six distances among the all four points are computed and are parts of the descriptor. The total descriptor is only 30 values per conformer: four partial charge values, four alogP values, four donor flags, four acceptor flags, four positive charge flags, four negative charge flags, and six distances. While on the surface, this small descriptor space seems implausibly simple; the canonical sorting of the points by charge actually encodes a hidden layer of information. For example, if the alogP value of feature point one is hydrophobic in nature, then we can conclude that the most negatively charged portion of the molecule is also hydrophobic.

To consider that multiple flexible conformers of a molecule exist in solution, we must compute the FEPOPS descriptor for multiple possible conformers. We systematically rotate all flexible bonds at 90° rotations, ruling out conformers with van der Waals clashes, and keep up to 1024 conformers per molecule. Interestingly, because of the fuzziness of the FEPOPS representation, very few conformers are needed to represent most of the available conformer space. Again, clustering is used to obtain a smaller number of conformers. This time, \( k \)-medoids clustering is used. In contrast to a \( k \)-means centroid—the geometric center of a cluster—a medoid is the actual data point that is closest to the center of the cluster. In other words, the medoid is a value from the original data set, whereas a centroid is not. We chose \( k \)-medoids to cluster conformers to ensure that we kept “real” conformers and not artificial averages that may be impossible to achieve in solution due to steric constraints. After benchmarking several \( k \) values for \( k \)-medoids clustering, it was determined that \( k = 7 \) best fulfilled the trade-off between covering the full FEPOPS conformer space while minimizing calculation time and descriptor storage. Figure 10.2 exemplifies for two different compounds how multiple conformers of flexible ligands are first turned into four centroids using \( k \)-means, and the resulting “FEPOPS conformers” are then clustered by \( k \)-medoids to a maximum of seven medoid conformers. Of course, more rigid molecules will have fewer than seven FEPOPS conformers. The seven irregular tetrahedrons are what form the basis of similarity searching.

Prior to descriptor calculation, we compute 3D coordinates, protonate the compound, set ionizable groups at pH 7.4, and compute partial charges and atomic logP values (Figure 10.3). In the original implementation of FEPOPS in 2003, as part of a Novartis postdoctoral project, molecular preparation was carried out by a...
A mixture of Pipeline Pilot (Accelrys) components and custom C code modified from a prior code written by Meir Glick et al. for multiscale docking [3]. Conformers were generated by systematic rotation of bonds using $90^\circ$ torsions. Note that multiple angles were benchmarked, from 10 to $120^\circ$. Angles less than $90^\circ$ were deemed redundant in the fuzzy representation, whereas angles greater than $90^\circ$ did not

Figure 10.2  Generation of feature point pharmacophores. The entire computed conformational space of two different compounds is shown at the left (above and below dashed line). For each conformer, atoms have been clustered by $k$-means and the centroids are shown (middle panel). From this set, $k$-medoids is used to pick a maximum of seven FEPOPS conformers. The centroids, properties, and distances are stored.

Figure 10.3  Assignment of atomic properties to centroids. Atoms of nevirapine were clustered in four $k$-means clusters and the centroids retained. For each centroid, the sum of partial charges ($q$) and sum of atomic logP ($p$) are assigned. In addition, a binary flag for presence of donor (HD) and acceptor (HA), positive charge ($P$), and negative charge ($N$) are assigned. Finally, the six distances between the centroids are measured.
fully capture the conformer space [5]. For compounds with five or fewer rotatable bonds rotated at the four 90° torsions, the allowance for 1024 conformers meant that all possible conformers are computed (4^5). For compounds with more than five rotatable bonds, a diverse sampling was obtained. Both k-means and k-medoids calculations were performed in R statistical packages.

In the second implementation of FEPOPS in 2006, all aspects of the code in Pipeline Pilot, C, shell scripts, and R clustering were replaced by a single Java executable (GGA, Moscow, Russia). Molecular 3D coordinates, protonation states, and partial charges were precomputed using Omega software (OpenEye), although the code is not dependent on any one molecular modeling package. At the time of writing this chapter – a decade after the original implementation – a database of FEPOPS descriptors is still precomputed regularly for the Novartis screening library. For convenience, FEPOPS similarity searches are still wrapped in a Pipeline Pilot protocol, which enables the generation of color-coded reports where atoms of query compounds and hit compounds are colored according to each of the four k-clusters. Similarity searching the precomputed Novartis FEPOPS binary database of >1 million compounds takes <30 s.

10.4 Scaling and Correlations

FEPOPS similarity is computed using Pearson correlation among all medoid conformers of any two molecules, and the highest correlation is kept. This assumes that the highest correlating conformers represent the best alignment between the two molecules. FEPOPS has also been used with a “bioactive conformation” of the query molecule taken from an X-ray structure. Interestingly, we see little advantage of using a single X-ray conformation in place of all medoids in terms of recall of active molecules from a database (see Section 10.6). Pearson was chosen due to the continuous nature of the charge, log P, and distance portions of the descriptor stored in a matrix format. Correlations tend to be highly positive on average due to the enforced charge-based sorting; thus, similarity cutoffs of r > 0.88 are typical for calling hits.

The original implementation of FEPOPS implemented a “mean center scaling” of the descriptor values by mean centering (offsetting values such that their sum is zero) and then dividing by a factor so that the variance of the scaled data is equal to one. Scaling factors were recorded after precomputing the entire FEPOPS database, for example, for the Novartis compound archive so that new query molecules not in the collection could be scaled in accordance with the search database. Because the descriptor itself contains few values, the correlations are susceptible to outliers in one of the feature properties. Thus, in FEPOPS v.2, we replaced mean center scaling with a Softmax scaling function to squash the range of values into a more linear range, which improved similarity performance in internal benchmarking studies (Mariusz Milik, Ben Cornett, unpublished).
10.5 Defining Scaffold Hopping

As discussed in Section 10.1, it is imperative to define a chemotype or scaffold in order to quantify the performance of a method that claims to scaffold hop. At the time that FEPOPS descriptors were being developed in 2003, few publications were available that defined scaffold hopping in a quantitative way. We began by surveying various definitions of scaffolds. Work around chemical graph representation had been published by Bemis and Murcko [5], a portion of which, resulting in a Pipeline Pilot implementation, is known as Murcko Assemblies (Figure 10.4). These assemblies are simply the connected sets of ring systems in a compound, inclusive of ring heteroatoms but without R groups. We can further reduce the scaffold by pruning more information, such as removing heteroatoms from the ring system, removing aromaticity, removing ring carbon count, and even disconnecting the rings. Many of these scaffold representations were available in an approach known as meqi (molecular equivalence indices) published by Xu and Johnson [6], which provided a deterministic license plate style representation for each reduced structure to enable comparison between molecules. Reduced topological representations such as meqi could thus be used to judge whether a query compound in a similarity search was a different scaffold from the hits that it recalled from a database.

In addition to using meqi for judging scaffold hopping, we also implemented a system of counts called SCINS (scaffold identification and naming systems) [7]. SCINS uses counts of the following features: (i) number of chain assemblies, (ii) number of chains, (iii) number of rings, (iv) number of ring assemblies, (v) number of bridge bonds, (vi) number of ring assemblies consisting of exactly one ring, (vii) number of ring assemblies consisting of exactly two rings, (viii) number of ring assemblies consisting of ≥3 rings, (ix) number of macrocycles, (x) binned length of shortest chain, (xi) binned length of second shortest chain, (xii) binned length of third shortest chain, (xiii) binned length of fourth shortest chain. If the binned lengths of chains exist, they are used; otherwise they are set to zero. The binned lengths follow the following scheme: one atom = 1, two atoms = 2, three or four atoms = 3, five or six atoms = 4, ≥7 atoms = 7).

![Figure 10.4](image-url)  
Figure 10.4  Topological reduction to quantify scaffold and ring system hopping. The reduced ring system is a visual representation of the SCINS descriptor.
SCINS can easily be computed on the fly, for example, in Pipeline Pilot. They provide a very liberal definition of scaffold, useful for objectively understanding if a hit compound is significantly different from a query compound, without relying on other 2D descriptors to make the comparison.

Oftentimes, the performance of similarity methods is judged using active recall, where a seed set of known actives (against a protein target, for example) are placed into a larger database or presumed inactives. The search method ranks the entire database and the rank of the seed actives is assessed using AUC (area under the curve) and computing ROC (receiver operator characteristic) scores to understand the enrichment over random across the rankings. Realistically, as a small number of compounds will be tested from a ranked database, many researchers simply look at enrichment for actives at the highest percentile ranking (e.g., top 1% or better). Recall that picking 1% of the compounds from a library of one million compounds still leaves 10,000 compounds to test. More often, it is advisable to look at recall in the top 0.01–0.1% to better reflect the needed enrichment levels to impact real-world projects.

In contrast to active recall tests—in which 2D similarity methods perform surprisingly well—scaffold-hopping tests require the cheminformatician to highly rank bioactive compounds with different chemotypes from a query molecule. In fact, a perfect scaffold-hopping method would have a near random enrichment of scaffolds while recalling active molecules. Stated another way, if a similarity method shows no preference for scaffolds yet can identify bioactive molecules from a query compound, the method is performing perfectly.

10.6
FEPOPS in Similarity Searching and Scaffold Hopping

Two compounds are similar according to FEPOPS if the feature point values correlate highly. For example, the FEPOPS descriptor for protein kinase C (PKC) inhibitor balanol has high FEPOPS similarity to ATP (Figure 10.5), an interesting example of a natural product correlating with a common endogenous cofactor.

FEPOPS was benchmarked for both active recall performance and scaffold hopping by pitting it against three industry standard 2D chemical descriptors: FCFP_4 Functional Class Fingerprints (Accelrys, San Diego, CA), MDL Keys (166 publicly available set, MDL Information Systems, Inc., San Leandro, CA), and DAYLIGHT fingerprints (Daylight Chemical Information Systems, Inc., Mission Viejo, CA). The FCFPs are based on an extension of the Morgan algorithm. Similarity was calculated using the Tanimoto coefficient for all 2D methods. In cases where more than one query compound was input, the maximum Tanimoto of database compounds to any of the probes was used to rank the database. In addition to the 2D approaches, we benchmarked pharmacophore distance triplets (PDTs), a commercial 3D method (Sybyl 6.9, Tripos, Inc., St. Louis, MO). PDTs contain three atoms and the binned distances among them, which are stored as a single bit in a binary fingerprint. Bits are assigned a 1 when a triplet occurs during
conformational enumeration. PDT does not require hand-built pharmacophores, and thus allows for a reasonable comparison to FEPOPS, which is deterministic and free of manual intervention. Details on our PDT parameterization were published previously [4].

Because the original FEPOPS benchmarking predated the existence of quality public compound-target data sources such as ChEMBL [8], test sets for similarity searching and scaffold hopping were taken from the commercial MDDR (MDL Drug Database Report). Sets were selected using MDDR classification of database records by biological activities or targets. From the entire MDDR set, a diverse subset of 30 000 compounds was selected as an “inactive” background set, which we called MDDRd. We found that similarity searches using either MDDR or MDDRd yielded similar active recall percentages for each descriptor. MDDR activity classes were chosen to represent major drug target classes: enzymes, G protein-coupled receptors (GPCRs), nuclear receptors (NRs), and ligand-gated ion channels (LGICs). The benchmarking data set consisted of 642 PTGS2/COX-2 and 597 gag-pol HIV reverse transcriptase enzyme inhibitors, 788 HTR3A/5-HT3A LGIC modulators, DRD2/D2 dopamine receptor GPCR modulators (151 agonists

Figure 10.5 The FEPOPS alignment of the protein kinase C inhibitor balanol and the kinase cofactor ATP. Atoms in clusters 1, 2, 3, and 4 are shown in their respective circles for the highest correlated FEPOPS conformers for these two molecules. The FEPOPS alignment reflects the crystal structure conformations of the two compounds complexed with PKC [9].
and 431 antagonists), and 331 RARA NR antagonists. For testing, the target class actives were seeded into the background MDDRds data set to assess the ability of the similarity methods to enrich for actives. A random molecule was chosen from each data set to be used as a query compound, except for the DRD2 and RARA cases where the endogenous ligand was used as the query: dopamine and retinoic acid, respectively (Table 10.1). In addition to MDDR sets, we selected four sets from Novartis HTS campaigns (Table 10.2). The HTS targets were GHSR growth hormone secretagogue receptor (GPCR), PSEN1/PSEN2 gamma-secretase (enzyme), MMP13 matrix metalloprotease-13 (enzyme), and a point mutation variant of ABL1 kinase (T315I). For background to the HTS-derived sets, 57 017 randomly selected inactive compounds from our HTS-plated compound library were chosen. For the HTS-derived sets, we used multiple probes (5, 10, or 20) rather than a single query compound, except for the case of ABL1, where the Novartis drug imatinib was the query compound.

The number of compounds that can be tested following a similarity search is limited in practice. For this reason, we selected a strict cutoff of the top 1.0% to assess the enrichments. It is clear that majority of the actives do not need to be recalled in order to return a sizable percentage of actives, active scaffolds, and ring systems (Table 10.1). In this test set, FEPOPS is ideal for capturing the largest number of scaffolds within a reasonable number of “cherry picks.”

As an alternative to our quantitative scaffold-hopping metrics, we also measured the average Tanimoto similarity (using MDL Keys) between the probe molecules and the seed actives in each data set (Table 10.3). An ideal scaffold-hopping method

<table>
<thead>
<tr>
<th>Data set</th>
<th>Percentage of recall in top 1%</th>
<th>FEPOPS</th>
<th>FCFP_4</th>
<th>MDL keys</th>
<th>Daylight</th>
<th>PDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTGS2</td>
<td>Actives</td>
<td>5.8</td>
<td>7.2</td>
<td>1.6</td>
<td>6.1</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>22</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Ring systems</td>
<td>39</td>
<td>17</td>
<td>22</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>HTR3A</td>
<td>Actives</td>
<td>5.3</td>
<td>8.2</td>
<td>6.1</td>
<td>6.0</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>17</td>
<td>14</td>
<td>18</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Ring systems</td>
<td>27</td>
<td>20</td>
<td>23</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>Gag-pol</td>
<td>Actives</td>
<td>3.5</td>
<td>1.3</td>
<td>1.5</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>9.0</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Ring systems</td>
<td>17</td>
<td>14</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>DRD2</td>
<td>Actives</td>
<td>1.8</td>
<td>1</td>
<td>0.1</td>
<td>0.7</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Ring systems</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>RARA</td>
<td>Actives</td>
<td>24</td>
<td>11</td>
<td>13.3</td>
<td>10.4</td>
<td>8.6</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>37</td>
<td>33</td>
<td>30</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Ring systems</td>
<td>60</td>
<td>40</td>
<td>47</td>
<td>13</td>
<td>47</td>
</tr>
<tr>
<td>Average</td>
<td>Actives</td>
<td>8</td>
<td>5.74</td>
<td>4.52</td>
<td>4.88</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>18.6</td>
<td>13.4</td>
<td>12.2</td>
<td>10.8</td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td>Ring systems</td>
<td>30.6</td>
<td>19</td>
<td>20</td>
<td>9.6</td>
<td>23.4</td>
</tr>
</tbody>
</table>
Table 10.2 Recovery of validated HTS actives, clusters, scaffolds, and ring systems by similarity searching.

<table>
<thead>
<tr>
<th>HTS target</th>
<th>Number of probes</th>
<th>Method</th>
<th>Percentage of total actives retrieved in top N%</th>
<th>Percentage of active scaffold classes in top N%</th>
<th>Percentage of active ring system classes in top N%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N = 1</td>
<td>N = 0.5</td>
<td>N = 0.25</td>
</tr>
<tr>
<td>GHSR4^a</td>
<td>5</td>
<td>FEPOPS</td>
<td>13</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>FCEP_4</td>
<td>11</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>FEPOPS</td>
<td>14</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>FCEP_4</td>
<td>15</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>PSENI1^a</td>
<td>5</td>
<td>FEPOPS</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>FCEP_4</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>FEPOPS</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>FCEP_4</td>
<td>9</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>MMP13^b</td>
<td>5</td>
<td>FEPOPS</td>
<td>24</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>FCEP_4</td>
<td>23</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>FEPOPS</td>
<td>27</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>FCEP_4</td>
<td>25</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>T18J ABL1^b</td>
<td>htratNB</td>
<td>FEPOPS</td>
<td>24</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>htratNB</td>
<td>FCEP_4</td>
<td>22</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

---

^a GHSR data set includes 438 actives, 26 structural clusters, 151 reduced scaffolds, and 45 reduced ring systems.

^b PSENI1 data set includes 331 actives, 17 structural clusters, 115 reduced scaffolds, and 47 reduced ring systems.

^c MMP13 data set includes 157 actives, 16 structural clusters, 43 reduced scaffolds, and 23 reduced ring systems.

^d ABL1 data set includes 182 actives, 13 structural clusters, 53 reduced scaffolds, and 23 reduced ring systems.
should find a diverse set of active compounds rather than enriching for actives that are consistently similar to the probe molecule. While the average Tanimoto similarity between probes and data set actives was 0.40, FEPOPS identified hits that were on average 0.44 similar to the probe molecules, whereas FCFP_4 enriched for compounds with an average similarity to probes of 0.52. This finding complements the scaffold-hopping performance determinations.

In cases where more than one probe molecule is known a priori, there is an opportunity to use the probes as an ensemble for database querying. We compared FEPOPS with FCFP_4 using either 5 or 20 probes for four targets, where all database molecules were ranked by maximum correlation to any single probe. Percentile rank cutoffs of 0.25, 0.5, and 1% were examined. With the exception of the case of PSEN1 as a target and 20 probe compounds as input, FEPOPS was comparatively better at scaffold hopping and ring system hopping (Table 10.2).

Putting the MDDR and HTS-derived data set results together, we find that FEPOPS allows for a good trade-off between active recall and chemical dissimilarity to probe molecules (Figure 10.6). PDT, an automated pharmacophore triplet approach, also showed good dissimilarity to query molecules but fewer actives recalled.

### 10.7 Alternative Alignment

The sorting of FEPOPS by partial charges – while canonicalizing the representation – also introduces artificial constraints on alignment between two molecules, and likely increases the false negative rate in large-scale screens. Nettles et al. introduced the CombiAlign version of FEPOPS [9], which attempted all combina-

<table>
<thead>
<tr>
<th>Target</th>
<th>Ligand</th>
<th>Average similarity of probe to all target actives</th>
<th>Average similarity of probe to actives recalled by FEPOPS (in top 1%)</th>
<th>Average similarity of probe to actives recalled by FCFP_4 (in top 1%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COX_2</td>
<td>SC-558</td>
<td>0.31</td>
<td>0.33</td>
<td>0.41</td>
</tr>
<tr>
<td>5-HT3A</td>
<td>Extreg 194, 584</td>
<td>0.45</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td>HIV-RT</td>
<td>Extreg 236, 942</td>
<td>0.39</td>
<td>0.45</td>
<td>0.66</td>
</tr>
<tr>
<td>D2</td>
<td>Dopamine</td>
<td>0.26</td>
<td>0.24</td>
<td>0.32</td>
</tr>
<tr>
<td>RAR</td>
<td>Retinoic acid</td>
<td>0.49</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>mABL</td>
<td>Gleevec</td>
<td>0.51</td>
<td>0.54</td>
<td>0.56</td>
</tr>
</tbody>
</table>

μ = 0.40  μ = 0.44  μ = 0.52

*Note*: Tanimoto similarities were calculated using public MDL Keys.
torial alignments of any two FEPOPS conformers being compared. In principle, this approach should capture more of the true hits (false negatives in traditional approach) but also introduce more false positives. A small, but significant boost in active recall was achieved using CombiAlign versus charge-based feature point sorting. At the same time, the number of comparisons between two molecules jumps 24-fold, from an average 49 (seven query medoids × seven test medoids) to 1176 (24 combinations of four features × seven medoids × seven test medoids). Notably, using only “shape” to sort the feature points performed worse than using charges or all possible orientations.

In practice, the choice of whether to use CombiAlign or the original charge-sorted FEPOPS is often determined by whether the probe has any notable partial charges that help to orient it; for compounds whose k-medoid clusters contain mostly neutral atoms, a charge-based canonicalization introduces potential randomness into alignments. For example, in the original FEPOPS approach, if feature point one, the most negative atom cluster, and feature point four, the most positive atom cluster, have relatively similar sum-of-partial-charge values, then the assignment of atoms to atom clusters 1 through 4 becomes arbitrary. This limits the accuracy of the method during similarity searching. In these cases, CombiAlign provides a better ratio of true positives to false positives.

It is worth noting that after years of using FEPOPS it has been observed that the “best” alignment of any two bioactive molecules is often not the conformation that those two molecules take on when bound to the same target active site. For example, we observed for the PTGS2 inhibitor SC-558 that active recall in similarity searching was stronger for the fully flexible probe than for the bioactive conformation alone [4]. In fact, Nettles et al. titled a manuscript that succinctly describes this phenomenon: “flexible 3D pharmacophores as descriptors of
For illustrative purposes, imagine cocrystal structures of some target X bound to active compounds Y and Z. The conformers of Y and Z found in their crystal structures may not correlate better in pharmacophore space than other possible conformations of Y and Z. Table 10.4 shows the results of a small study of the active and scaffold recall of aldose reductase inhibitors, using either their cocrystal structure conformations or fully flexible conformations to derive the FEPOPS descriptors. Accuracy was measured using the AUC for either active or active scaffold recall from a background data set. With the exception of zopolrestat, in most of the cases, the flexible probe set performed better than the bioactive conformers. On some level, this observation may reflect the real-life flexibility of protein residues and bound compounds: in solution, the compound and protein breathe in a dynamic way. Many suboptimal conformers of compounds Y and Z precede and follow the actual binding event to target X, that is, “subbinding” conformations. Because subbinding conformations of Y and Z may be quite similar as well, the restricting probe molecule conformations to crystal structure poses alone may not be the best strategy. We believe that the FEPOOPS representation in particular is poised to find higher correlated subbinding conformations due to its fuzzy placement of features.

### 10.8 In Silico Target Prediction

Another lesser-known application of similarity searching algorithms is to search databases of biologically annotated molecules with an orphan probe molecule in order to infer a target. Biologically interesting SMUTS (Small Molecules with Unknown TargetS) can arise from a number of ways. Natural products are often found to possess interesting biological properties in assays, but their targets remain

### Table 10.4  Recall of MDDR aldose reductase inhibitors (73) and aldose reductase inhibitor scaffolds (15) with bioactive (X-ray) and flexible pharmacophores.

<table>
<thead>
<tr>
<th>Pharmacophore probe</th>
<th>Rotatable bonds</th>
<th>Probe conformation</th>
<th>ROC AUC$^a$ actives</th>
<th>ROC AUC$^a$ scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tolrestat</td>
<td>6</td>
<td>X-ray</td>
<td>0.84</td>
<td>0.93</td>
</tr>
<tr>
<td>Tolrestat</td>
<td>6</td>
<td>Flexible</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td>Alrestatin</td>
<td>2</td>
<td>X-ray</td>
<td>0.85</td>
<td>0.94</td>
</tr>
<tr>
<td>Alrestatin</td>
<td>2</td>
<td>Flexible</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>IDD384</td>
<td>8</td>
<td>X-ray</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>IDD384</td>
<td>8</td>
<td>Flexible</td>
<td>0.81</td>
<td>0.93</td>
</tr>
<tr>
<td>Zenarestat</td>
<td>4</td>
<td>X-ray</td>
<td>0.60</td>
<td>0.70</td>
</tr>
<tr>
<td>Zenarestat</td>
<td>4</td>
<td>Flexible</td>
<td>0.79</td>
<td>0.91</td>
</tr>
<tr>
<td>Zopolrestat</td>
<td>5</td>
<td>X-ray</td>
<td>0.75</td>
<td>0.92</td>
</tr>
<tr>
<td>Zopolrestat</td>
<td>5</td>
<td>Flexible</td>
<td>0.71</td>
<td>0.89</td>
</tr>
</tbody>
</table>

$^a$ Receiver operating characteristic, area under the curve.
unknown. In modern times, the largest source of SMUTS is phenotypic screening, or compound screening with cellular or in vivo conditions that more accurately reflect a disease-relevant state. In some cases, high-throughput screens in cell-based systems can lead to many thousands of hits. Only a very limited number can be mechanistically pursued due to the overhead of target identification. For large lists of SMUTS, computational approaches to predict targets are clearly preferred. For example, 2D fingerprints of compounds in large compound-target bioactivity databases have been used to train statistical models that enable rapid target prediction and annotation of SMUT lists [10]. Such approaches have been dubbed in silico target fishing [11]. At the current time of writing, ligand-based Bayesian models for >4700 human proteins have been created at Novartis based on integrated internal, commercial, and public compound-target data sources, representing a respectable percentage of the human genome. However, 2D in silico target fishing, like all QSAR approaches, is limited by domain applicability issues: the less the SMUT probe molecule resembles the chemogenomics search database, the lower the prediction accuracy is. In these cases, a 3D or 4D similarity descriptor may provide a way forward for in silico target fishing in lieu of successful 2D approaches. For example, in head-to-head comparisons of FEPOPS and ECFP_6 (Pipeline Pilot, Accelrys) in target prediction, ECFP6 outperformed FEPOPS on average [12]. However, in cases where the SMUT probe had Tanimoto similarity less than 0.8 (MDL public keys) to the next nearest neighbor in the reference chemogenomics database, FEPOPS ability to predict small-molecule targets outperformed ECFP_6. These data suggest that initial 2D similarity assessment of the SMUT probe to the reference database should provide guidance as to whether 2D or 3D in silico target fishing is more likely to succeed. It is typical for natural products to have low 2D similarity to bioactivity databases containing mostly synthetic compounds from medicinal chemistry programs. We suggest that pharmacophore approaches such as FEPOPS may be more useful for natural product target identification.

10.9 Chemical Space Uniqueness

FEPOPS has been used in several studies that attempt to understand how different chemical descriptors behave in comparison to one another in clustering bioactive compounds. Schuffenhauer et al. ask the basic question of whether clustering compounds by chemical structure has any biological meaning [7]. They used Pareto analysis to optimize two concurrent objectives, cluster size and spread in biological activities among a compound set. A mix of descriptors was used for three pharmacology profiling experiments (in vitro safety profiling panel, kinase panel, NCI60 cell proliferation). The authors concluded that no classification method is overall superior, although in cases of a large number of clusters, scaffold driven approaches provide more biologically meaningful clusters, whereas in cases of a smaller number of clusters, fingerprinting methods work better. In many cases,
FEPOPS is less useful for biologically relevant clustering; however, in the case of safety profiling data, FEPOPS outperformed all other methods when there were fewer than 20 clusters. This suggests that FEPOPS is able to capture more global similarity among a large number of diverse compounds.

Hert et al. used FEPOPS and other descriptors to create drug class trees and quantify their relationships [13]. They asked whether a chemical organization of targets could reflect biology. To do so, they quantified the differences between cheminformatics and bioinformatics approaches to target similarity. For example, does the similarity of compounds known to modulate proteins yield a similar set of relationships as the default approach of comparing proteins by amino acid sequence? The authors used a mix of 2D descriptors (Unity, MDL keys, ECFP_4, FCFP_4, Daylight) and 3D approaches (CATS and FEPOPS). Similarity between 249 proteins via cheminformatics was determined by the similarity ensemble approach (SEA) and Bayesian model similarity. Overall, FEPOPS similarities between target-set probes showed low correlation to the similarity matrices generated by all other chemical descriptors, suggesting that FEPOPS has a very different perspective on what makes any two compounds similar. Interestingly, it was also concluded that the proteins found to be similar by cheminformatics similarity matrices were completely uncorrelated to similarity matrices generated by PSI_Blast, or sequence similarity.

To more generally understand the “descriptor space” of commonly used chemical descriptors, Bender et al. used PCA (principal components analysis) on lists of database molecules ranked by similarity to a large number of probes [1]. The authors picked 11 activity classes from MDDR and 37 different descriptors, selecting 10 random compounds per activity class as probes and tracking the number of actives recalled by each descriptor in the top 5% of the ranked compounds. Once again, FEPOPS appeared to exist in a unique spot in descriptor space, far away from the 2D methods. Its closest neighbor in the PCA plot was the TGD descriptor, typed graph distances with atom typing (donor, acceptor polar, anion, cation, hydrophobe) as implemented in MOE (CCG). Also nearby was the SCINS descriptor, a reduced ring system described earlier in this chapter. In terms of active recall, FEPOPS showed lower performance than the 2D methods, but (in contrast to our original FEPOPS paper [4]) looked similar to other methods in terms of reduced ring system hopping.

10.10
Perspective on FEPOPS’ 10 Year Anniversary

FEPOPS has contributed many novel bioactive molecules to our lead discovery and lead optimization programs on an yearly basis and was used on well over a 100 projects within the first few years it was developed. For example, in a prospective scaffold hop [14], a potent CCR3 antagonist from a competitor company was used as a query molecule to search for similar compounds (Figure 10.7). The top-ranking database molecule is shown at the right, a compound that when tested had
an IC₅₀ = 6 nM for CCR3. The drug-like size of the probe molecule and cationic charge made it ideal for FEPOPS searching. Indeed, a number of chemotypes were identified by FEPOPS on the basis of a limited number of probes.

Scaffold hopping approaches are abundant now. At the time that FEPOPS was developed in 2003, we were unable to find any papers that quantified scaffold-hopping performance. We created a means of quantifying scaffold hopping and reduced ring system hopping and compared the new descriptor to best in class 2D descriptors. After a decade of use, it is clear that our most used 2D methods (ECFPs) outperform FEPOPS in simple active recall measures, but that FEPOPS consistently returns strikingly different database compounds and is near-random in its chemotype preference. The near-random chemotype preference reflects FEPOPS’ uniqueness in descriptor space and also suggests that it is the most powerful when used side by side with a good 2D method for real pharmaceutical research projects. When coupled with a reasonable cherry picking capacity and the throughput of plate-based assays, FEPOPS is a useful means of scaffold hopping from known probes to help drive forward projects with new chemical matter.

References


American Chemical Society, 124 (10), 2337–2344.


11
Three-Dimensional Scaffold Replacement Methods

Nathan Brown

11.1
Introduction

The wider availability in recent decades of three-dimensional (3D) structures of small molecules, whether modeled, small-molecule crystal structures, or protein–ligand cocrystals, has enabled these data to be applied to many challenges in computer-aided molecular design. Not the least of these is the application of scaffold hopping using the available 3D coordinate data to identify potential scaffold replacements, while also maintaining the geometries of substituent functional groups in the geometric orientations that are observed to be important, for example, for binding.

A number of methods exist with the objective of replacing scaffolds with alternative, yet appropriate, scaffolds using defined exit vectors. This chapter will consider three of the leading methods published in the literature: SHOP, ReCore, and BROOD. However, the workflow for each of these methods is conceptually similar, but with different algorithms; Section 11.2 defines the generic workflow for 3D scaffold replacement.

11.2
Generic Three-Dimensional Scaffold Replacement Workflow

The 3D scaffold replacement workflow applied in the approaches defined later in this chapter follows a very similar workflow. This section details each of the components of the generic workflow for 3D scaffold replacement. Figure 11.1 defines the context of each of these components within this workflow.

11.2.1
Molecule Databases

Scaffold replacement tools require a database of previously defined fragment scaffolds with which to evaluate possible replacements. There are a number of databases that have been used in the recent studies of scaffold replacement...
methodologies, including small-molecule crystal structures, exemplified medicinal chemistry space, commercially available compounds, and virtual scaffold libraries. Each of the methods described in this chapter requires a 3D structure of each fragment to be considered, but some tools do permit the generation of these structures on the fly as a part of the individual workflows.

CSD Structures: The Crystal Structure Database is available from the Cambridge Crystallographic Data Centre and represents a library of almost 600,000 small-molecule X-ray crystal structures [1]. Since the structures have been experimentally determined, some favor this library. However, the subsequent fragment generation step removes these moieties from their parent molecule potentially resulting in 3D conformations for these fragments that may not be appropriate for these approaches. Furthermore, the CSD contains many different types of molecule, many of which may not contain medicinal chemistry-relevant fragments of interest and these should be filtered after fragment generation.
ChEMBL Structures: Recently, a number of large databases of medicinal chemistry-relevant molecules have been published. One of the larger of these collections, abstracted from the medicinal chemistry literature, is ChEMBL [2]. Currently, ChEMBL contains almost 1.3 million unique compounds representing an enviable library. Furthermore, since the ChEMBL database has been abstracted from medicinal chemistry-relevant journals, it is more likely that the resultant fragments will be of interest as scaffold replacements.

Commercial Structures: Commercially available compounds number in many millions, with eMolecules [3] currently containing more than 5 million unique compounds that one may purchase from a variety of compound vendors. This database represents a significant quantity over and above those represented in the CSD and ChEMBL, but these compounds tend to have been synthesized from off-the-shelf building blocks that, when fragmented, merge together to provide a relatively smaller fragment space than one would expect. However, more recently, companies are focusing on molecular diversity and this will tend toward increasing this fragment diversity in the future.

Virtual Structures: The last database to be considered here is virtual scaffold libraries: libraries of enumerated scaffolds based on certain principles, such as particular molecular framework or through optimization. One example of a virtual scaffold library comes from the work of Pitt et al. [4] in which the authors enumerated all likely fused heteroaromatic ring systems that had no literature precedent at the time of publication.

The above four databases represent exemplars of the types of databases that may be applied in 3D scaffold replacement, but many others are available. The challenge of selection of potential replacement fragments depends on the following criteria: quality of the 3D conformations, the novelty of the replacement scaffolds, and synthetic precedence in the literature. Any or all of the above databases may be included in a 3D scaffold replacement campaign, depending on the requirements of the project.

11.2.2
Fragment Generation and Filtering

For each of the databases discussed in Section 11.2.1, a fragmentation step is performed to identify potential scaffolds, including their exemplified exit vector patterns or simply potential exit vectors.

REtrosynthetic Combinatorial Analysis Procedure (RECAP) is the most commonly used fragmentation from the literature [5], but any fragmentation procedure may be used. RECAP proceeds by applying a set of 11 rules for cleaving bonds in a given molecule. However, often all possible combinations of cuts are used to fragment the molecules while also retaining the cut points and their relative orientations of their torsion angles.

Once the fragments have been generated, it is prudent to remove any fragments that contain undesirable properties such as size and complexity. These properties
should be nonadditive, since other properties (such as ClogP) may be modulated favorably once placed in the context of the original substituents. For efficient search and retrieval, the fragments are often stored in a relational database with an appropriate chemistry-aware cartridge.

11.2.3 Fragment Replacement Search and Scoring

Once the fragments have been generated and filtered, they may be applied in identifying appropriate replacements of a scaffold in the context of the original substituents. Geometric alignment of the exit vectors is desired as part of the alignment optimization and these are scored using a number of methods based on the level to which features and geometric locations are retained in the replacement.

Once a replacement has been identified and calculated to be feasible, it is often desirable as a postalignment operation to model potential conformational changes. Once the substituents have been appended to the replacement fragment, it is important to identify whether the substituents will remain in the preferred geometric locations. One can achieve this by performing a conformational search for favorable, low-energy conformations and determining from those conformations the root-mean-square deviation (RMSD) in ångströms of the substituents or simply the dihedral angles of the substituent points. However, the latter approach, although more rapid, may not consider possible intramolecular interactions that could lead to more significant changes in distal geometries.

11.3 SHOP: Scaffold HOPping by GRID-Based Similarity Searches

SHOP was first described in a paper by Ahlström et al. [6] and is now marketed commercially by Molecular Discovery Ltd. The first report of SHOP in the literature used only GRID molecular interaction fields (MIFs) from which to calculate the scores of the possible replacement scaffolds [7].

In 2009, SHOP was extended to include four different methods to calculate the score of the potential replacement: geometry (SHOP-geom), GRID MIFs (SHOP-GRID), shape (SHOP-Shape), and molecular fingerprints (SHOP-finger) [8,9]. SHOP-geom uses two descriptors: the first calculated as the geometric distance between the connection points and the dihedral angle of the vector between the two anchor points and their connected atoms. SHOP-GRID uses small-molecule probes to simulate chemical groups representing hydrogen bond acceptor and donor characteristics, hydrophobicity, and electrostatic interactions. The MIF for the possible replacement scaffold is then generated and compared with the reference scaffold, with the attachment points being encoded as hydrogen for the purpose of the MIF calculation and comparison. SHOP-shape uses the positive energy MIF points using the N1 probe (amide nitrogen probe with one hydrogen atom attached to it, the H-bond donor). The frequency of occurrence of distances
between the anchor point and the selected positive MIF points is then encoded in a vector. SHOP-finger uses a fingerprint representation of atoms, which are encoded as hydrophobic, hydrogen bond donor, hydrogen bond acceptor, formally positively charged, and formally negatively charged. The geometric distances from each anchor point to each atom position are then calculated and the distances encoded as Gaussians with an alpha factor of 0.5.

The four calculated SHOP descriptors are then used to compare the reference scaffold with each possible replacement scaffold in the fragment database. All combinations of different anchor points are considered in the comparison with the highest similarity comparison retained in the final ranked list of similarities.

The SHOP method was applied retrospectively to thrombin, HIV-1 protease, and neuraminidase in a retrospective known scaffold recovery experiment and in a prospective, using docking as prospective validation, experiment to identify novel scaffold replacement. The results of these experiments were very positive and subsequent papers have also shown that SHOP is effective at identifying promising novel replacement scaffolds [10].

11.4 ReCore

ReCore was first published in 2007 and describes a methodology that has been embodied in a software program commercially available from BioSolveIT as part of the LeadIT software suite [11].

The fragment library used in ReCore uses the RECAP set of retrosynthetic rules to fragment the CSD. The authors stated their intent to concentrate on experimentally observed conformations, but highlight that the system can use any set of molecules from which to derive fragments. All possible combinations of cuts were used to generate the fragments. Once generated, the fragments undergo a filtering process to remove any undesirable fragments. The size of the fragments and the number of cuts in the fragments are applied early since they are additive properties.

One of the key strengths of ReCore is the indexing system applied to index each of the fragments for rapid search and retrieval using R-trees for storage. A k-nearest neighbor search is then applied to find appropriate replacements.

ReCore has been applied in studies to identify potential scaffold hops for the following targets: HIV protease, HIV reverse transcriptase, and thyrotropin-releasing hormone. For each target, the system was successful in rediscovering extant ligands or identifying highly similar analog ligands demonstrating the utility of the system.

11.5 BROOD

OpenEye Scientific Software offers BROOD for bioisosteric replacements using a range of different bioisosteric similarity principles to effectively identify bioisos-
teres that are chemically interesting and likely to be synthetically tractable. The four bioisosteric similarity methods BROOD uses are as follows:

1) Overlap of shape, atom types, and attachment geometry
2) Overlap of shape, electrostatics, and attachment geometry
3) Preservation of attachment geometry, not including chemical properties
4) Close analogs to the query fragment

These four bioisosteric similarity methods are called color, elect, struc, and queryAnalog in the BROOD system, respectively. The elect hitlist is computationally much more expensive than the other three to calculate and therefore this is not calculated automatically, but the user may choose to select this comparison should the query require. These four methods are each calculated separately and separate hit lists generated for consideration by the user. BROOD itself does not conduct any extensive conformational analyses, but the software permits the user to retrieve only cyclic replacements that tends toward smaller, more rigid replacements.

BROOD comes with a fragmented database that is generated from approximately 12 million compounds that are commercially available, which reduces some of the potential issues with synthetic tractability. The database provided by OpenEye contains 90 000 simple or 600 000 more complex unique fragments. These fragments are additionally filtered to remove toxic, reactive, and large fragments.

Recently, BROOD has been released with a Graphical User Interface (GUI) called vBROOD intended for both molecular modelers and medicinal chemists.

BROOD has been retrospectively validated with known bioisosteric pairs, while the system has also been validated to retrieve more generic structures that are less structurally similar to queries. BROOD has been used to identify possible bioisosteric replacements to design novel GABAA α3 modulators by subsequent generation of a virtual library using the possible replacements.

11.6 Conclusions

Both structural databases and databases of exemplified and virtual fragments continue to grow at a rapid rate. With the more information available on the potential replacement scaffolds and experimentally observed or modeled conformations, it is anticipated that the reliability of these 3D scaffold replacement tools described in this chapter will improve. It is quite often intuitive when looking at the results from these structure search methods, but as with many computational modeling methods, it is important to use a variety of validation approaches to ensure

---

that the resultant molecules do indeed exhibit the functional group geometries desirous for making important interactions with protein targets of interest.

Acknowledgment

N.B. is funded by Cancer Research UK Grant No. C309/A8274.

References

12
Spherical Harmonic Molecular Surfaces (ParaSurf and ParaFit)

David W. Ritchie and Violeta I. Pérez-Nueno

12.1
Introduction

This chapter presents a completely scaffold-free method for comparing the 3D surface shape and other surface properties of molecules using spherical harmonic (SH) functions. It begins by summarizing the basic theory behind the SH shape matching approach. It then introduces the ParaSurf and ParaFit programs. Finally, it presents a case study of using the SH representation to cluster and classify multiple entry-blocking ligands for the CCR5 cell surface coreceptor, and it demonstrates how the notion of consensus surface shapes may be used as scaffold-hopping VS queries to explore how different ligands might distribute themselves within the CCR5 extracellular pocket.

12.2
Spherical Harmonic Surfaces

SHs play an important role in physics and engineering applications because they often appear as the solutions to certain differential and integral equations [1]. Consequently, they can be used to represent and model many natural phenomena, ranging from electrostatics and electromagnetism to quantum theories of matter. The complex SH functions, $Y_{lm}(\theta, \phi)$, are often expressed as a product of normalized Legendre polynomials, $P_{lm}(\cos \theta)$, and complex circular functions:

$$Y_{lm}(\theta, \phi) = \left[ \frac{(2l+1)(l-m)!}{4\pi(l+m)!} \right]^{1/2} P_{lm}(\cos \theta) e^{im\phi},$$

(12.1)

where $\theta$ and $\phi$ are the usual spherical coordinates (Figure 12.1), and the subscripts $l$ and $m$ label different functions subject to the restriction that $-l \leq m \leq +l$. SH functions are orthonormal in the sense that the overlap integral of a product of two
SHs is unity or zero according to whether the two functions are the same or different, respectively, that is,

$$\int_0^{2\pi} \int_0^\pi Y_{lm}(\theta, \phi)^* Y_{l'm'}(\theta, \phi) \sin \theta \, d\theta \, d\phi = \delta_{ll'} \delta_{mm'},$$  \hspace{1cm} (12.2)

where $$\delta_{ij}$$ is the Kronecker delta, and the asterisk denotes complex conjugation. Here, it is convenient to use the related real SH functions, often denoted as $$y_{lm}(\theta, \phi)$$, which may be found by taking linear combinations of the complex functions:

$$y_{lm}(\theta, \phi) = \begin{cases} (Y_{lm}(\theta, \phi) + Y_{lm}(\theta, \phi)^*)/\sqrt{2}, & \text{if } m > 0, \\ Y_{10}(\theta, \phi), & \text{if } m = 0, \\ -i(Y_{lm}(\theta, \phi) - Y_{lm}(\theta, \phi)^*)/\sqrt{2}, & \text{if } m < 0, \end{cases}$$  \hspace{1cm} (12.3)

where $$\bar{m} = -m$$. The real SH functions are also orthonormal. Figure 12.1 shows the shapes of the real SHs up to order $$l = 2$$.

For present purposes, the orthonormality property means that SHs may be used as basis functions, or “shape building blocks,” to encode and reproduce the detailed shape of a molecular surface. For example, if a molecule is placed at the coordinate origin, the radial distance from the origin to the molecular surface may be encoded as a sum of real SHs to order $$L$$, using

$$r(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta, \phi),$$  \hspace{1cm} (12.4)
where $a_{lm}$ are the expansion coefficients. An expansion to order $L$ will have a total of $(L + 1)^2$ coefficients. Using the orthogonality of the basis functions, it is straightforward to show that each coefficient may be calculated as

$$a_{lm} = \int_0^{2\pi} \int_0^\pi r(\theta, \phi) y_{lm}(\theta, \phi) \sin \theta \, d\theta \, d\phi.$$  \hspace{1cm} (12.5)

In other words, each coefficient is uniquely determined by the degree of overlap between the shape to be represented and the corresponding basis function.

One disadvantage of the SH representation, however, is that it can only faithfully encode surface shapes that are single valued, or star-like, with respect to radial rays from the chosen origin. It cannot represent multivalued surfaces. One way to circumvent this problem, at least partially, is to calculate and encode the solvent-accessible surface (SAS) of a molecule. This is the surface that is swept out by rolling a spherical water molecule over the van der Waals surface of the molecule. For many small molecules, the SAS surface is often more star-like than the van der Waals surface. Figure 12.2 shows the SAS surface of lorazepam calculated to various SH polynomial expansion orders. Using higher order polynomials obviously gives a more accurate representation of the molecular surface. However, we find that truncating the expansion at $L = 6$ (i.e., using just $7^2 = 49$ coefficients to represent the shape of each molecule) encodes sufficient detail for VS.

12.3 Rotating Spherical Polar Fourier Expansions

It can be shown that the SH functions of each order $l$ transform among themselves under a general Euler rotation according to [2,3]

$$\hat{R}(\alpha, \beta, \gamma) Y_{lm}(\theta, \phi) = \sum_m D_{lm/m}^{(l)}(\alpha, \beta, \gamma) Y_{lm'}(\theta, \phi),$$  \hspace{1cm} (12.6)

where $\hat{R}(\alpha, \beta, \gamma)$ represents a rotation operator expressed in terms of the Euler angle parameterization $(\alpha, \beta, \gamma)$. In other words, a rotated SH function can be calculated as a linear combination of unrotated functions of the same order. The rotation matrices, $D^{(l)}(\alpha, \beta, \gamma)$, are originally due to Wigner [4], although it seems...
he never published a full derivation. For complex SHs, the Wigner rotation matrices have the form

\[
D_{m'm''}^{(l)}(\alpha, \beta, \gamma) = e^{-im'\alpha}d_{m'm''}^{(l)}(\beta)e^{-im\gamma},
\]

(12.7)

where the real \(d_{m'm''}^{(l)}(\beta)\) functions are themselves closely related to the spherical harmonics [3]. They may be calculated efficiently using a variety of recursion formulae. As might be expected, it is also possible to take suitable linear combinations of the Wigner rotation matrices in order to obtain the corresponding real rotation matrices, \(R_{m'm}^{(l)}(\alpha, \beta, \gamma)\), for use with the real SHs [5]. However, the details are rather tedious, and are not given here. In any case, rather than rotating the SH basis functions, it is often more useful to work with a set of fixed SHs and to rotate instead the expansion coefficients. For example, once the real coefficients of a SH surface have been determined, the representation may be rotated into an arbitrary orientation rather efficiently using

\[
\hat{R}(\alpha, \beta, \gamma)r(\theta, \phi) = \sum_{l=0}^{l} \sum_{m=-l}^{l} \left( \sum_{m'=-l}^{l} R_{m'm''}^{(l)}(\alpha, \beta, \gamma) a_{lm} \right) y_{lm}(\theta, \phi).
\]

(12.8)

It is worth noting that for a pure rotation about the z-axis, this reduces to

\[
\hat{R}(\alpha, 0, 0)r(\theta, \phi) = \sum_{l=0}^{l} \sum_{m=-l}^{l} (a_{lm} \cos m\phi + a_{lm} \sin m\phi) y_{lm}(\theta, \phi).
\]

(12.9)

Therefore, for the best performance, and to sample rotational space as evenly as possible, we rotate the coefficients of a molecular surface in two stages. We first use Eq. (12.8) with \(\alpha = 0\) to calculate a set of rotated coefficient vectors using a near-regular distribution of pairs of \((\beta, \gamma)\) angular samples from an icosahedral tessellation of the sphere (Figure 12.1). We then calculate the final z rotation as needed using Eq. (12.9).

### 12.4 Spherical Harmonic Surface Shape Similarity

Having calculated the SH surface shape of a molecule, we can essentially discard the underlying covalent scaffold, and retain only the SH expansion coefficients as a kind of “pseudomolecule.” We can then calculate the similarity between a pair of such pseudomolecules (which are assumed to be centered at the coordinate origin) by searching for the rotation that minimizes the squared “Euclidean distance” between the corresponding pairs of SH surface shapes, \(r_A(\theta, \phi)\) and \(r_B(\theta, \phi)\):

\[
D_{\text{Euclidean}}^2 = \int_0^\pi \int_0^{2\pi} (r_A(\theta, \phi) - \hat{R}(\alpha, \beta, \gamma)r_B(\theta, \phi))^2 \sin \theta \, d\theta \, d\phi.
\]

(12.10)
Thanks to the orthonormality of the basis functions, this expression reduces to
\[ D_{\text{Euclidean}}^2 = |a|^2 + |b|^2 - 2a \cdot b', \]  
where \( b' \) represents the vector of rotated SH expansion coefficients of molecule B, and so on. This expression has units of \( \text{Å}^2 \) and it clearly depends on the relative size of the molecules being compared. However, it is worth noting that the first two terms are rotation-invariant. Only the final scalar product depends on the relative orientations of the two pseudomolecules. Therefore, when comparing multiple molecules, it is often convenient to use a normalized similarity score in which identical molecules give a score of unity. For example, dividing Eq. (12.11) by the sum of squared magnitudes of the SH shape vectors and rearranging it gives an expression that has the form of a Hodgkin similarity score [6]:
\[ S_{\text{Hodgkin}} = \frac{2a \cdot b'}{|a|^2 + |b|^2} = 1 - \frac{D_{\text{Euclidean}}^2}{|a|^2 + |b|^2}. \]  
In a similar spirit, we can also define a Carbo-like [7] and a Tanimoto-like [8] similarity score using
\[ S_{\text{Carbo}} = \frac{a \cdot b'}{|a| \cdot |b|} \]  
and
\[ S_{\text{Tanimoto}} = \frac{a \cdot b'}{|a|^2 + |b|^2 - a \cdot b'}. \]
These will both be equal to unity when \( a = b' \).

### 12.5 Calculating Consensus Shapes and Center Molecules

An additional benefit of the SH representation is that it is straightforward to calculate the average shape, or “consensus shape,” of a given group of molecules, simply by summing the coefficients of the individual SH surfaces. The overall principle is illustrated in Figure 12.3. For example, given \( N \) coefficient vectors, the consensus shape, \( \tilde{r}(\theta, \phi) \), may be calculated as
\[ \tilde{r}(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \left( \frac{1}{N} \sum_{k=1}^{N} a_{lm}^k \right) y_{lm}(\theta, \phi). \]  
Of course, the individual surfaces must first be rotated into an optimal mutual superposition before calculating the average of their SH coefficients. In practice, we do this iteratively. For example, by first calculating Tanimoto similarity scores for all pairs in the group, the two most similar SH shapes are selected and superposed to form a seed consensus shape. Then, the remaining molecules are
superposed onto the seed consensus, and a new consensus shape is calculated as the average of the SH coefficients of the superposed shapes. The individual consensus members are then superposed again onto the consensus average, and the process is iterated until convergence. We normally find that this procedure converges in just two or three iterations. The result is a consensus pseudomolecule and a list of superposed molecules, each individually rotated into the orientation of the consensus.

Obviously, calculating consensus shapes results in some smoothing and loss of detail compared to the individual molecular shapes. However, this can be considered a desirable property because it provides an unbiased way to combine the most significant features of a related group of molecules. For example, the consensus shape representation can be used to capture the essential 3D shape features of several known high-affinity ligands and to encode them in the form of a single representative pseudomolecule that may be used as a VS query. Furthermore, once a SH consensus shape has been calculated for a group of functionally related molecules, it is straightforward to identify a “center molecule,” that is, the real molecule whose SH surface is closest to that of the consensus shape. For drug targets with multiple known active molecules, calculating the consensus shape and real center molecule provides a rational way to select a single VS query molecule.

12.6
The ParaSurf and ParaFit Programs

The above-mentioned techniques have been implemented in the ParaSurf and ParaFit programs, available from Cepos Insilico. For each molecule, ParaSurf reads the quantum mechanical electron density from MOPAC [9] or VAMP [10], which it contours using a “marching cubes” algorithm with a threshold of $2 \times 10^{-4} \text{e/Å}^3$ (which is approximately equivalent to the molecular SAS) [11]. The SH

surface shape coefficients are then calculated by numerical integration over this surface. In any multivalued or nonstar-like regions of the surface, the greatest radial surface distance is taken. Thus, the resulting surface may be considered as a kind of “shrink-wrap” envelope [12] that surrounds the true surface. ParaSurf also calculates certain key local surface properties (namely, electrostatic potential, ionization energy, electron affinity, and polarizability), which may also be encoded as SH expansions [13]. However, we find that the single most important quantity for VS is surface shape [14]. Hence we do not consider these additional properties here.

The ParaFit program reads and compares the SH coefficients calculated by ParaSurf. The two programs communicate using an in-house extension to the “SDF” (Structure Description File) file format. ParaFit can process an arbitrary number of molecules loaded from one or more multimolecule SDF files. The only limitation is the total amount of computer memory available. Typically, ParaFit treats the first molecule as a “query” molecule, and all subsequent molecules as a “database” against which the query is searched. ParaFit provides three main calculation modes. In the default “fitting” mode, ParaFit superposes one or more moving database molecules onto a single fixed query molecule. For convenience, all results are presented in the coordinate frame of the query. The program can also perform all-versus-all superpositions in which each molecule is superposed in turn onto all others. In this “matrix” mode, a table of distance scores is written out in a format suitable for subsequent clustering analysis, for example. In addition to superposing molecules, ParaFit may also be used to calculate the consensus shape of a group of molecules and to calculate the corresponding real center molecule.

ParaFit superposes molecules using a brute-force rotational search over the three Euler rotation angles. The Tanimoto shape similarity score is the default scoring function, although any combination of properties using any of the above-mentioned similarity measures may also be used. Conceptually, each moving molecule is rotated with respect to the fixed query molecule, and the Euler rotation that gives the greatest similarity score is recorded. This is essentially a Fourier correlation search in Euler angle coordinates. However, because good superpositions may be achieved using only low-order harmonic expansions, we find that it is not necessary to use fast Fourier transform (FFT) techniques to accelerate the calculation unless \( L \geq 16 \).

In addition to using low-order correlation searches, ParaFit’s superposition calculations are accelerated in two further ways. The first technique exploits the fact that harmonic expansions to order \( L \) can have no more than \( L^2 \) local maxima. Hence, ParaFit initially uses relatively large angular search steps of around 8° to cover the search space. In order to sample angular space evenly and efficiently, these angular samples are generated from the vertices of an icosahedral tessellation of the sphere, as illustrated in Figure 12.1. This gives around 30% fewer sample points than a naïve equiangular grid [5]. Once the approximate location of maximum similarity has been identified, it is then refined using a localized grid search in steps of 2°. Both angular step sizes may be adjusted by the user.
The second acceleration technique is used when comparing multiple molecules. Rather than rotating each database molecule in turn, it is more efficient to rotate the SH expansions of only the query molecule and to compare these against each fixed database molecule. Thus, relatively expensive SH rotations are applied to just one rather than N molecules. Once the optimal rotations have been found, the database molecules are rotated using the inverse of the corresponding query rotations. Using these techniques, a pair of molecules may be superposed in around 0.05 seconds on a 1.8 GHz Pentium Xeon processor, and computation times may be further reduced by a factor of about 5 when searching a database.

### 12.7 Using Consensus Shapes to Probe the CCR5 Extracellular Pocket

This section briefly summarizes a SH-based VS study of human immunodeficiency virus (HIV) “entry blockers” that we performed using ParaFit [15–17]. HIV infection is initiated by fusion of the virus with the target cell through binding of the viral gp120 protein with the CD4 cell surface receptor protein and the CCR5 and CXCR4 coreceptors. There is therefore considerable interest in developing new ligands that can modulate the conformations of these coreceptors and hence block virus–cell fusion. We have studied both CCR5 and CXCR4 as VS targets, although here we focus on CCR5 because it has a large number of known antagonists from some 16 different scaffold families. Thus, it demonstrates our SH-based approach from a scaffold-hopping point of view. For that study, we collected a total of 424 CCR5 antagonist inhibitors from the literature, and constructed a database of about 4700 decoy molecules with similar physicochemical properties to the known actives. However, several earlier computational studies had indicated that different CCR5 ligands bind in fundamentally different ways within the CCR5 extracellular pocket [16,18–21]. Hence, we knew a priori that it would be very unlikely that a single VS query structure would be able to retrieve all of the actives from our database. This was quickly confirmed by the rather disappointing VS results that we obtained when we screened our database using the high-affinity TAK-779 ligand as the query [15].

In order to construct more effective VS queries, and to probe how the different scaffold families might distribute themselves within the CCR5 pocket, we clustered the 424 known CCR5 inhibitors using Ward’s hierarchical clustering of chemical (topological) fingerprints as implemented in the JKlustor module of JChem.\(^2\) We used Kelley’s method [22] to calculate an optimal number of 16 clusters, which we then reduced by visual inspection to just 10 clusters that group together compounds belonging to similar scaffold families. We then calculated the consensus shapes of these 10 clusters using ParaFit, as described earlier. However,
because we believed that the CCR5 ligands would bind in somewhat fewer than this number of subsites of the receptor pocket, we used ParaFit again to compute all-against-all rotational superpositions of the 10 fingerprint-defined consensus shapes. This produced a shape-based similarity matrix that we reclUSTERed to identify clusters of similar consensus shapes. These were again superposed and averaged to give a total of four “super-consensus” (SC) surface shapes that we labeled A–D, as shown in Figure 12.3. Cluster A contained mostly TAK derivatives and anilide piperidine N-oxides (87 compounds), cluster B contained guanylhydrazone and 4-hydroxy-piperidine derivatives (69 compounds), cluster C contained various pentacyclics, pyrrolidine piperidines and other piperidines, the oxopryrrolidine-carboxamides, diphenyl ureas, diketopiperazines, propanediamines, and the SCH and AMD derivatives (184 compounds), and cluster D mainly contained other propanediamines, piperidines, tropanes, and the phenylcyclohexylamines (84 compounds).

To measure the ability of these scaffold-free SC shapes to identify known binders, we performed VS using our CCR5 ligand database, and we analyzed the screening utility of each SC query using receiver operator characteristic (ROC) plots. We use the area under the curve (AUC) of each ROC plot to give a single numerical measure of the ability of each consensus shape to recognize known actives with similar shapes. An AUC of 1.0 represents perfect VS performance (i.e., all of the actives are retrieved before any of the inactives), whereas an AUC of 0.5 corresponds to a random selection of actives and inactives. However, because early retrieval of actives is import in VS, it can also be useful to consider the shape of the ROC curve.

Overall, we found that SC query C gave the best VS performance with an AUC of 0.91. This is perhaps not surprising because that cluster includes the three most active compounds in the database, and also a large number of other actives with similar shapes to the 4-piperidine derivatives, the SCH derivatives, and the 1,3,4-trisubstituted pyrrolidine piperidine derivatives. The A query also performed well with an AUC of 0.79, and the D query performed reasonably well with an AUC of 0.63. The ROC plot for query B gave a rather low AUC of 0.41, although this query gives good sensitivity and selectivity in the first percentages of the database screened. However, if the two small B and D queries are merged, the early performance profile is lost and the screening performance becomes essentially random, with an overall AUC of 0.51. Thus, despite the small populations of these two groups, their members have significantly different overall shapes, and they should be treated as two distinct structural groups for VS purposes. We applied a similar exercise with other combinations of SC clusters, and found similar but less dramatic reductions in AUC compared to the unmerged clusters. We therefore concluded that the CCR5 antagonists might be grouped into no less than four main SC families [16].

To explore whether the four SC pseudomolecules are sterically feasible within the CCR5 extracellular pocket, we calculated the volumetric union of the ligands from each SC cluster and docked the resulting pseudomolecules into our model-built structure of CCR5 [16] using the Hex rigid-body docking program [23]. The docking
results suggested that the SC shapes may be docked to three overlapping regions within the CCR5 extracellular pocket, with pseudomolecule A docking to the first subsite delimited by transmembrane (TM) helices 1, 2, 3, and 7, pseudomolecule C docking to a second subsite (TMs 3, 5, and 6), and with both pseudomolecules B and D docking to a third region (TMs 3, 6, and 7), which overlaps the first two subsites.

From a ligand-based VS point of view, these docking results suggested that it could be fruitful to treat each CCR5 subsite as a distinct VS target. Therefore, we tested this hypothesis by recalculating the ROC curves with clusters B, C, and D treated as inactives with respect to query A, and similarly by treating clusters A, B, and D as inactives with respect to query C, and finally with clusters A and C treated as inactives with respect to the B and D queries. This gave new AUC values of 0.83, 0.96, and 0.85 for the three target subsites. Given that SCs A and C already gave good ROC curves, reassigning the B and D members as inactives only marginally improved the corresponding AUCs. On the other hand, treating the A and C members as inactives for the B and D queries significantly improved the final AUC value. Overall, the improved AUC values suggest that it can be useful to consider different receptor subsites as distinct VS targets, and further supports the notion that CCR5 has three ligand-binding subsites within its extracellular pocket. More generally, from our experience of this computational scaffold-hopping experiment, we suggest that for VS targets for which multiple actives are known, it can be a useful strategy to perform scaffold-free shape clustering to predict whether the target might have multiple subsites, and to consider using one or more consensus shapes as effective VS queries.

12.8 Conclusions

This chapter has shown how SH functions may be used to represent the 3D shapes of small molecules. In particular, the SH representation provides a convenient scaffold-free way to compare molecular shapes and to calculate the consensus shapes of groups of molecules. We have briefly described one example application of using the SH approach to cluster multiple scaffold families of CCR5 antagonists in order to give a relatively small number of super-consensus shapes, which could then be used to model how different ligand families might distribute themselves within the CCR5 extracellular pocket. While this largely ligand-based computational experiment seemed to produce a plausible model of different receptor–ligand binding modes for CCR5, until we obtain a crystal structure of the receptor, such a model must remain speculative. Nonetheless, the ease with which the SH shapes of the diverse CCR5 ligand scaffolds could be calculated, compared, and clustered demonstrates the utility of using the SH representation to test and support scaffold-hopping hypotheses about ligand–target binding without requiring a crystallographic structure of the target.
References


8 Stewart, J.J.P. (1990) QCPE #455, Quantum Chemistry Program Exchange, Bloomville, IA.


inhibitors: partial and full antagonists. 


The XED Force Field and Spark

Martin Slater and Andy Vinter

13.1 Pharmacological Similarity — More than Just Chemical Structure

This chapter is a part of the larger section on “shape-based methods for scaffold hopping”. The term “shape based” puts us firmly into three dimensions but implies that “shape” is thought of in terms of bonds and atoms that we use to describe the chemical structure of a molecule. Plainly, an active ligand bound to its protein-binding site must have a 3D shape that is compatible with that site, but it is not the shape alone that determines the biological effect (Box 13.1).

A ligand must present binding features that are correctly placed to interact maximally with the corresponding protein groups that make up the binding site. These binding features take the form of van der Waals contacts and electrostatic attractions and repulsions such as hydrogen bonds, ion pairs, and $\pi-\pi$ interactions.

We think of this shape compatibility in terms of the final bound state, but ligands may move through a series of conformations before reaching the binding site. Therefore, the binding pattern must also be flexible enough to allow the ligand to enter the protein at the correct “gate” and make its way toward the final binding state. Again, shape alone cannot be enough to provide this nonequilibrium directional incentive, but the electrostatic forces and close-packed sequential van der Waals attractions can.

For example, class A G protein-coupled receptors (GPCRs) seem to pull in their charged ligands by strong ion-pair attraction to an oppositely charged residue set deep within the receptor. This must be accompanied by changes in shape as the ligand burrows in toward its final pose, implying that new shapes must be generated and be conformationally allowed. The strong, long-range electrostatic interactions represent the “engine” driving the binding process, while the weak, short-range attractive van der Waals forces are the engine’s fine tuners and settlers.

We must conclude that the similarity of chemical structure alone is not enough to define pharmacological similarity if scaffold hopping is to be diverse enough to rid us of the “chemotype trap.” One way to include the potential bonding features onto the structure of a ligand is to map the molecular electrostatic potential close to the van der Waals surface. This involves probing all parts of the molecule with a
nominal cation to find nucleophilic regions, then again with a nominal anion to define complementary electrophilic areas, and finally with a neutral van der Waals atom to find the short-range “sticky” areas. Other probes can be used to highlight further properties. The large amount of data arising from probing in this way presents computational difficulties. A more compact representation is to distill the regions down to just their extrema: we term these “field points” (Box 13.2).

The pattern that a set of field points generates from a given 3D molecule is characteristic of that molecule. Many potential ligand conformations can be processed and their field patterns compared without worrying about the structural features that generated them. Field points can even be generated for proteins,

### Box 13.1

<table>
<thead>
<tr>
<th>Chemical Structure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2</td>
<td>Histamine H2 receptor antagonist</td>
</tr>
<tr>
<td>Thrombin</td>
<td>PDE3 enzyme inhibitor</td>
</tr>
<tr>
<td>PDE3</td>
<td>11βHSD inhibitor</td>
</tr>
<tr>
<td>11βHSD</td>
<td></td>
</tr>
</tbody>
</table>

Shape alone is not enough to define pharmacological property. Examples of diverse chemotypes that act at the same biological site are illustrated. At the histamine H2 receptor, both Cimetidine and Ranitidine act as antagonists. cAMP is the natural substrate for the PDE3 enzyme that is inhibited by the pyridazinone. Two well-known inhibitors of Thrombin are shown, only one of which is peptidic. The tetrazole is just one example of many chemotypically diverse inhibitors of the enzyme 11βHSD.
allowing the detection of possible binding and allosteric sites, of high and low solvation areas and remote electrostatic variations that would attract diverse ligands.

It follows that if two molecules of comparable volume, regardless of their structural detail, can generate a similar field pattern, then they are likely to have similar pharmacological properties (Box 13.3). It is sobering to realize that field point patterns vary not only with substitution but also with conformation (Box 13.4).

Field points are the distillation of the 3D molecular electrostatic potential (MEP) down to the positive points of maximum electrophilicity (red) and the negative points of maximum nucleophilicity (blue). Also plotted are hydrophobic points where coulombic influence is at a minimum (orange) and van der Waals points (yellow) indicating important surface “stickiness.” The sphere radius of each field point reflects its potential strength of interaction. Field points render field information computationally viable.
When molecules encounter each other, they do not see the structure as a set of strung atoms; they see molecular fields generated by the structure. In their bioactive conformations, cAMP, the natural substrate for phosphodiesterase 3, and a known pyridazinone inhibitor of PDE3, generate unexpectedly similar field patterns despite being chemically very different. When these two field patterns are optimally matched, the likely reason for inhibition is revealed.

Box 13.4

The MEPs change considerably with conformation.
Using the field point concept, scaffold hopping is carried out by comparing the field point patterns of tens or hundreds of conformations of an active molecule with a similar number of a new molecule until a good match is found. The new molecule may be one of millions in a database, each of which is held in many feasible conformations.

13.2 Improving the Generation of Valid Molecular Fields

Ligand fields reflect what a protein or any adjacent molecule “sees” and reacts to in terms of shape and electrostatics. The creation of a useful field pattern depends on the credibility and distribution of the generated atomic charges. Quantum mechanical (QM) methods are best placed for this task but cannot be fast enough to process the vast number of conformations or sizes of some candidates in an acceptable time. The only alternative is molecular mechanics (MM).

However, most commonly used MM force fields contain a serious scientific flaw when working on this scale. They place single atomic charges at atom centers, regardless of how they are generated. This is tantamount to collapsing the atomic electrons into the nucleus, thus denying any discrimination of the electron distribution or the nuclear charge from the electronic charges [1]. Coulombic interactions over long distances absorb this approximation but at short distances when probing for field points, unacceptable losses of definition occur, particularly across aromatic and H-bond regions.

In response to this ACC (atom centered charge) approximation, a new MM force field was begun in 1990 [2] to overcome many of the published failures to use fields as property indicators (Box 13.5). The main motive behind the development of a new MM methodology was to improve the quality of the molecular electrostatic and dispersive potentials around a molecular surface but also resulted in a robust means of structural optimization that overcame many shortcomings of other MM packages. Once quality field generation was established, molecular field similarity needed for real chemotype hopping (see Box 13.3) could be developed [3].

Box 13.5

The XED model reflects intuitive molecular properties (left) where the ACC model (right) fails.
The main experimental tool that is available to validate the position and pattern of field points is provided by IsoStar, “a library of information about nonbonded interactions,” [4] that utilizes the extensive Cambridge Crystallographic Database. (Box 13.6) We have made extensive use of this in parameterizing the force field.

**13.3 The eXtended Electron Distribution (XED) Force Field**

The eXtended electron distribution (XED) force field has been under continuous development since 1990 and has been successfully applied in academic and commercial drug discovery for over ten years, particularly as a reliable generator of molecular fields and as a means of optimizing large and small structures. The XED MM methods resolve a number of hitherto unresolved approximations by commercial MM force fields (see below) including π–π interactions, electron polarizability, and atom asymmetry.

The XED force field distributes electrons around the nucleus in a comparable way to quantum orbital constructs, but the form of the electron distributions for each π atom is empirically derived using a data set of experimentally determined conformational energy differences as validation (Box 13.7). Earlier attempts to use QM yielded complex patterns that were found not to perform better than the empirically derived ones.
A maximum of five XED constructs (termed “xed”) are used to describe the charge environment of an atom. Each of these has its own parameters for length, angle, and charge. Saturated atoms (e.g., Csp3) and sigma bond overlaps have no associated xeds. For all other \(\pi\)-orbital-containing atoms, standard charges are allocated to each xed on a given atom that sum to the allocated nuclear charge. We now have a positively charged nucleus neutralized by the surrounding negatively charged xeds.

For example, neutral divalent oxygen (Osp3) has a nuclear charge of +4, two 0.3 Å \(z(\pi)\)-xed at −1.0 each, two 0.3 Å \(y\)-xed at −0.5 each, one 0.3 Å +x-xed at −0.7, and one 0.2 Å x-xed at −0.3. Sigma and \(\pi\) partial charges are calculated from the Huckel theory and a bond dipole equalization scheme and are distributed onto each neutral xed-bearing atom, sigma charge to the nucleus and \(\pi\) charge to the xeds. Now the atom has incorporated partial charge and is polarizable when subjected to optimization, as the xeds can move in response to applied electric fields. One downside of the method is that there are roughly as many added xeds as there are.

---

**Box 13.7**

Atom centered charges for benzene and acetone exemplify the lack of detail necessary for sensible electrostatic calculations. Poor MM-based optimizations and MEPs result.

The XED constructs are polarizable and allow for nuclear charge. The torsional barriers for conjugated systems are controlled by the \(\pi-\pi\) XED out-of-plane potential.

---

A maximum of five XED constructs (termed “xed”) are used to describe the charge environment of an atom. Each of these has its own parameters for length, angle, and charge. Saturated atoms (e.g., Csp3) and sigma bond overlaps have no associated xeds. For all other \(\pi\)-orbital-containing atoms, standard charges are allocated to each xed on a given atom that sum to the allocated nuclear charge. We now have a positively charged nucleus neutralized by the surrounding negatively charged xeds.

For example, neutral divalent oxygen (Osp3) has a nuclear charge of +4, two 0.3 Å \(z(\pi)\)-xed at −1.0 each, two 0.3 Å \(y\)-xed at −0.5 each, one 0.3 Å +x-xed at −0.7, and one 0.2 Å x-xed at −0.3. Sigma and \(\pi\) partial charges are calculated from the Huckel theory and a bond dipole equalization scheme and are distributed onto each neutral xed-bearing atom, sigma charge to the nucleus and \(\pi\) charge to the xeds. Now the atom has incorporated partial charge and is polarizable when subjected to optimization, as the xeds can move in response to applied electric fields. One downside of the method is that there are roughly as many added xeds as there are.
hydrogens on a molecule resulting in slower energy minimisations than conventional force fields.

The main features of the force field are its general applicability and simplification compared with others. This is due mainly to the dependence of conformation on the alignment of z (\(\pi\)) orbitals that govern out-of-plane bending, conjugation, and cross-term electrostatic interactions, all of which have multiple entries in the conventional force fields. The number of atom types and entries for bond and torsional parameters is greatly reduced and the inclusion of 1–3 electrostatic interactions allows the orbital asymmetry of many atoms to be handled. For example, the high conversion barrier of methyl acetate and the anomeric effect are correctly reproduced through consideration of 1–3 electrostatics, without requiring explicit parameterization.

The most important contribution made by adding xeds is that \(\pi\)–\(\pi\) interactions are realistically reproduced, whereas atom-centered charges give rise to a parallel intermolecular stack that is never seen in nature. (Box 13.8) Extensive validation has illustrated the improvement in electrostatic behavior of the XED system over other popular force fields [5] (Box 13.9).

13.4
The XED Force Field Applied to Scaffold Hopping in Spark

The XED force field underlies a range of commercially available computational chemistry software from Cresset. Box 13.10 briefly outlines some of the many applications spawned from field similarity.

Spark is the application used to generate bioisosteric substitutions. The following section describes in detail how the XED molecular mechanics works in Spark, with subsequent examples of its application in a drug discovery context.

13.5
How Spark Works

Spark requires the initial input of a molecular template structure in a defined 3D conformation. The choice of starting point is critical to this and many of the other field-based software applications, since fields are dependent on conformation. Ideally, the template is the biologically active conformation, or a close approximation to it. Typically, the template may be extracted from an X-ray protein–ligand complex, derived from a pharmacophore analysis of a chemical series, or computationally generated, for example, by using Cresset’s Forge software.

The XED force field is applied to generate the template’s electrostatic fields and field point patterns. A portion of the molecule is then selected and chosen as the bioisosteric fragment replacement site. An illustration of how this is done within the software’s graphical user interface is shown in Box 13.11 using the Cox2 inhibitor molecule rofecoxib as an example.
The desymmetrization of the ether oxygen atoms by XED addition accounts (singular) for the anomeric effect (left) and the large conformational energy difference (right).

The addition of XEDs to benzene reproduces the observed Bz/Bz associations (right). ACC docking (left) results in a parallel arrangement that is never seen experimentally. For each pair, best pose is left. To the right, Bz has been docked onto a central Bz from 120 points around a sphere of radius 10Å.
A database of 3D fragments is then searched to find suitable replacements for the selected region. The fragment databases are obtained by applying a fragmentation technique to the zinc database [6]. Each of the fragments has precomputed conformations and specified reconnection positions for rapid assessment in the 3D context of the template. The initial assessment of each fragment is purely geometric: it must possess attachment points in similar positions and at similar angles to the region of the template selected for replacement. This process is summarized in Box 13.12 using the rofecoxib example.

Fragments that pass the geometric screen are then merged into the template to create a new molecule. Care is taken that the merging process creates chemically
sensible results. In particular, the charge states of atoms near the merged region are reassessed, as for example, appending an aromatic group to a charged primary amine renders it nonbasic.

The resulting molecule is minimized using the XED force field, and field point descriptors are applied. The similarity of the new molecule to the original template is then assessed using the field descriptors. Crucially, what is assessed is the shape and electrostatic similarity of the entire final molecule to the original template molecule. Existing techniques have tended to focus on assessing the similarity of the replacement fragment to the region of the template selected for replacement. Doing so ignores the fact that the electrostatic properties of a fragment can strongly depend on what it is connected to: a phenyl group with an electron-withdrawing substituent can have a markedly different electrostatic potential to the same group with an electron-donating substituent.

Each new molecule is minimized, aligned, scored, and then ranked by its score in the output list. Thus, the output from the virtual experiment is a ranked set of molecules, which encompasses the most useful potentially bioisosteric replacements from thousands of possibilities. The algorithms employed have been highly optimized, so that a database of hundreds of thousands of fragments can be processed within minutes.
13.6
Application of Spark in Drug Discovery Scenarios

There are many uses of bioisosteric replacements in drug discovery. Scaffold hopping is a key example, often used for patent breaking or patent fortification. For patent breaking, the goal is generally to replace the main core, or scaffold, of the drug. Patent fortification is more likely to involve an analysis of the scaffold’s decoration in order to ensure that the patent covers as much of the relevant chemical space as possible. Series optimization is another area of potential use, particularly where a specific biologically active group presents a liability and thus an ideal replacement would be a diverse chemistry sharing the biological activity only.

The following section describes another example of this powerful technology in a little more detail.
P38 belongs to the protein kinase superfamily of enzymes and has a key role in the mediation of cellular responses to stress. This kinase regulates cytokine expression and is implicated in the pathogenesis of a number of autoimmune and inflammatory diseases. Inhibitors of P38 are of great interest as therapeutic agents that may combat such diseases. However, to avoid the various off-target activities stemming from interaction with the extended family, high selectivity is an important requirement.

The recently reported P38 inhibitor PH878667 has exquisite selectivity toward the target kinase. This high selectivity is in part explained using X-ray crystal data for those compounds in P38 complexes that are observed to undergo a relatively rare protein conformation at the hinge region. The hinge region is situated between the N and C lobes of the kinase and is a well-conserved interaction site within this protein kinase family. The site provides critically important H-bond donor and acceptor patterns required for the recognition of the adenine heterocycle of ATP.
In P38, the hinge sequence contains glycine, which imparts enough flexibility to the backbone to allow easy rotation of this residue, which flips this pattern from the usual H-bond acceptor to donor. This change in interaction pattern is beautifully exploited by compounds such as PH234234, which specifically interact with this and not the standard configuration, thus explaining the high selectivity (Box 13.13).

P38 is a particularly flexible kinase, evidenced by the many available X-ray crystal structures known for this protein. The conformation described above is one of many that it is known to adopt. A further example and a therapeutically relevant P38 protein conformation is the inactive “DFG-out” conformation. This form of the P38 protein is equivalent to that which the kinase c-abl adopts during inhibition by the anticancer drug Gleevec.

This conformation is the result of a rearrangement of a flexible loop, termed the activation loop, which displaces residues critical for phosphoryl transfer. The conformational change is sensitive to the phosphorylation state of the full loop and has a time dependency on its formation. Inhibitors of the DFG-out state are described as non-ATP competitive and characterized as having slow on and/or slow off rate kinetics. A fragment inhibitor, bound into the “DFG-out” form of P38, is shown in Box 13.14.

13.8 Creating New Molecules

A hypothetical inhibitor using interactions of both the hinge-flipped and DFG-out protein conformations could potentially combine selectivity with slow off rate kinetics and confer an advantage. Box 13.15 shows the overlay of the selective inhibitor (PDB: 3HUB) with the “DFG-out” fragment (PDB: 3K3I).
Box 13.14

P38 in “DFG-out” from with fragment inhibitor

predominant hinge conformation

Graphics from Pymol by Delano Scientific

Box 13.15

Overlay of selective inhibito and “DFG-out” fragment

predominant hinge conformation

Graphics from Pymol by Delano Scientific
Spark was used to computationally grow the DFG-out fragment to create new molecules that match the fields of the combined molecules. Since a protein structure did not exist for the combined protein conformations at the time of this study, the fields of the two ligands and close residues were analyzed to check their compatibility prior to beginning the experiment.

Box 13.16 shows the fields of the two inhibitors and confirms that they have similar and compatible electronic field patterns, particularly with respect to the face of the molecules interacting with two close aromatic residues from each of the two protein conformations. The experiment was run using a dummy atom placed at a suitable position on the DFG-out fragment as the replacement site. The output molecules were scored by field similarity to the combined inhibitors with an 80:20 weighting in favor of the selective ligand.

13.9 New Potential Inhibitors

Spark provided a number of interesting output molecules that appeared to be feasible as potential inhibitors. Example 2D structures of output molecules and their 3D field point patterns are shown in Boxes 13.17 and 13.18, respectively.

Subsequent to this experiment, a protein X-ray structure was deposited in the PDB that provided the first evidence confirming the existence of these two
### Box 13.17

**Output examples: 2D structure and field similarity**

<table>
<thead>
<tr>
<th>Structure</th>
<th>Rank</th>
<th>Sim</th>
<th>Structure</th>
<th>Rank</th>
<th>Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Structure 1" /></td>
<td>1</td>
<td>0.528</td>
<td><img src="image11" alt="Structure 11" /></td>
<td>11</td>
<td>0.492</td>
</tr>
<tr>
<td><img src="image4" alt="Structure 4" /></td>
<td>4</td>
<td>0.507</td>
<td><img src="image13" alt="Structure 13" /></td>
<td>13</td>
<td>0.489</td>
</tr>
<tr>
<td><img src="image6" alt="Structure 6" /></td>
<td>6</td>
<td>0.501</td>
<td><img src="image53" alt="Structure 53" /></td>
<td>53</td>
<td>0.476</td>
</tr>
</tbody>
</table>

### Box 13.18

**Output examples: 3D structure, fieldpoints, and rank**

- Fragment and reference
- Rank 4
- Rank 6
- Rank 11
- Rank 13
- Rank 53
conformations combined in a single protein complex, with an inhibitor remarkably similar to those provided by this study. Interestingly, the profile of this drug candidate was developed for use as an inhaled formulation for COPD. Box 13.19 shows the 2D and 3D structures of further examples from the virtual experiment compared with this published COPD drug candidate.

13.10 The Far-Reaching Consequences of Using Molecular Fields as Measures of Similarity

The use of three-dimensional molecular fields to compare properties and similarities has now become wide spread. It is reasonably well accepted that fields are the drivers of molecular recognition while structure is only the generator of the field and not what another molecule “sees” or “feels.” It is worth repeating that if two potential ligands with different structures can generate similar fields, the binding patterns of those ligands, and hence their biological actions, will be similar.

There are many examples of drugs with very different structures that act at the same biological site and elicit similar biological responses. In many cases, comparison of their fields reveals why this is so and has some far-reaching consequences for drug design and beyond. It suggests that focused libraries should be based on similar field patterns rather than structural similarities. The “Markush” principles for patenting totally miss this vital aspect of protection because they are, by definition, limited to structural analogs. However, to attempt to change these institutionalized activities will be not be productive in the short term.
Many drug targets are members of a cascade of proteins, each dependent on the other for controlling action, feedback, and specific outcome. In the past, chemists have attempted to hit two or more sites with one molecule made from combining the structural features from drugs already known to act on one of the sites. For many obvious reasons, this stratagem has not been wholly successful. However, if the fields of two drugs that hit different parts of a cascade are combined into a field template and that template is used to search for a single molecule with a similar field pattern, new and different structures may be found that hit both sites and the concept of polypharmacology can be realized.

The extension of field analysis into protein–protein interactions opens up a means of isolating the field patterns of strong binding sites between proteins that can be “field templated” and used to search for small molecules that alter these sites for improvement or disruption. The implication again is that fields transcend molecular architecture so that the peptide structure is no longer an impediment to drug design.

With the expansion of protein modeling in structural and field space, we have come full circle, back to the XED molecular mechanics force field as the computational means to deal with proteins as accurately and practically as possible. The XED force field provides a reliable scientific basis that has overcome many flaws in previous offerings and made it possible to generate valid field patterns that have proven scientific value and successful application to drug discovery. The final battle to overcome the understandable adherence to chemical structure as a means of property comparison has yet to be won but in the areas of patent coverage, preclinical prediction, polypharmacology, protein–protein interactions, and many other areas, molecular fields are beginning to play an increasingly important part.

Acknowledgments

We are indebted to Katriona Scoffin and Mark Mackey for their help in constructing this chapter.

References


14

Molecular Interaction Fingerprints

Didier Rognan and Jérémy Desaphy

14.1 Introduction

Three-dimensional (3D) structures of protein–ligand complexes provide crucial information to better understand molecular rules governing living cells and assist rational drug discovery. If analyzing a few structures at a graphic desktop is now common practice, mining and comparing a large array of protein–ligand complexes requires a simplification of the 3D information. Among the most useful simplification processes for analyzing protein–ligand interactions is the conversion of atomic coordinates into simpler fingerprints. Fingerprints are easy to generate, manipulate, compare, and therefore enable a systematic analysis of large data sets. They are largely used to describe and compare molecular objects (small molecular weight ligands [1], pharmacophores [2], proteins [3], protein–ligand binding sites [4]) and represent descriptors utilized by computer-aided drug design programs, notably in silico screening tools [5]. Computational chemists frequently manipulate these fingerprints in either ligand-based or structure-based approaches to drug design. It is however interesting to combine both protein-based and ligand-based information in a single descriptor focusing on either simple association or even better, true molecular interactions. This chapter reviews the current status of molecular interaction fingerprints focusing on both development and application.

14.2 Target-Annotated Ligand Fingerprints

Among the several thousand existing descriptors for ligand structures [6] are simple 1D or 2D fingerprints encoding the presence or absence of key functional atom types or groups. Such topological fingerprints are usually used to assess pairwise similarity among ligands or evaluate the chemical diversity of compound libraries. Assuming that similar structures share similar bioactivity, fingerprints have been frequently used to screen compound archives for similarity to reference ligands and therefore identify novel bioactive compounds [7]. Although successful
in many instances, ligand similarity does not fully mirror biological similarity (only in about 30% of cases, as estimated in a survey of high-throughput screening data at Abbott [8]). One reason for this discrepancy is that all ligand atoms are not in direct interaction with host proteins. Ligand similarity (or dissimilarity) may therefore be explained by atoms not responsible for binding to a target. This observation led to the design of ligand fingerprints focusing on protein interacting atoms [9]. The fingerprint remains similar to a true ligand fingerprint (linear, circular) but describes only the interacting atoms instead of the entire molecular graph (Figure 14.1).

14.2.1 Interacting Atom/Fragment Fingerprints

The first attempt to restrict ligand fingerprints to protein interacting features was reported by Crisman et al. [10] in the IASF (interaction annotated structural features) method that assigns precalculated interaction energy weights to ligand atoms in direct contact with the target. Starting from a protein–ligand X-ray structure, interaction energies are calculated on a per atom basis and summed over each ligand substructure encoded in a standard extended connectivity fingerprint [11]. Virtual screening of database compounds is simply done by searching structural features that are common to that of the reference(s), cumulating their interaction energy scores, and scoring hits by decreasing cumulative scores. In three out of five retrospective virtual screens, the IASF scoring scheme was shown to outperform pure structure-based docking and ligand-centric similarity searches [10]. The method could be made even simpler by just isolating the suite of interacting atoms (called interacting fragment or IF) and outputing the IF as a
A classical keyed fingerprint (166-bit public MACCS structural keys) [12]. A problem with these fingerprints is their sensitivity to complexity effects. Hence, IF-derived fingerprints have a low bit density (fraction of on-bits) that makes it difficult to distinguish compounds of variable size. Scaling bit positions based on the relative frequency of the corresponding interacting fragment among true actives was shown to improve the usability of IF fingerprints [13]. Another potential problem is the generation of artifactual keys (resulting from disjoint IFs) that would be absent from the full atom fingerprint. Atom-centered fragments in which every interacting atom is iteratively stored considering its surrounding atoms, avoids the above-described drawback of IF-derived fingerprints. They were shown to enhance early enrichment in true actives in fingerprint-based similarity search experiments [11].

### 14.2.2 Protein–Ligand Pharmacophores

Interactions being easy to detect from protein–ligand coordinates [14], pharmacophoric properties (e.g., H-bond acceptor or donor) can be stored as well. Due to the ever-increasing number of high-resolution protein–ligand X-ray structures of pharmaceutical interest [15], a large collection of 68,000 protein–ligand interaction pharmacophores has recently been described and used for either ligand profiling (finding the most likely targets of a given ligand) or as a fast and reliable alternative to protein–ligand docking [16]. Pharmacophores are usually described by a set of features and their topological relationships, but can also be converted into a vector registering feature types and interfeature distances in specific bins [17]. Such pharmacophore-based interacting fingerprints (Pharm-IF) were recently used as input to train a support vector machine to distinguish, with high enrichment factors, true actives from decoys [17]. Although not directly encoding protein–ligand interactions, the FLAP (Fingerprints for Ligands And Proteins) method [18] generates four-point pharmacophore fingerprints from either ligand structures or molecular interaction fields directed against ligand-binding sites. It therefore provides a unique framework for mapping ligands to targets. Pharmacophores present the advantage to be fuzzy enough to accommodate partial matches (e.g., only a subset of features are matched by ligand atoms) and do not require energy calculations, the Achille’s heel of molecular docking. Since the pharmacophore concept is quite simple, it can be applied to any kind of interaction, notably to the important field of protein–protein interactions (PPIs) [19,20] in order to quickly find PPI inhibitors by a structure-based approach [21].

### 14.3 Ligand-Annotated Target Fingerprints

Ligand-annotated target fingerprints rely on the opposite construction to that related above, which means that the fingerprint describes specific protein atoms or residues interacting with a known ligand (Figure 14.2).
This concept has been extensively used to convert protein–ligand 3D information into molecular interaction fingerprints, notably for the purpose of postprocessing docking poses. Hence, docking is still hampered by the relative inaccuracy of fast-scoring functions to discriminate correct from wrong poses and to rank ligands by decreasing binding free energy [22]. An alternative strategy to rigorous prediction of free energies is to postprocess docking poses to prioritize the most reliable ones irrespective of energetic criteria. Among the many possible approaches, the usage of protein–ligand interaction fingerprints (IFPs) has gained considerable popularity [23].

An IFP is a simple bit string registering, from atomic 3D coordinates, the presence or absence of well-defined interactions (apolar, aromatic, hydrogen bonds, salt bridges, metal coordination) between a ligand and a fixed set of amino acids (or protein atoms) lining its binding site. Interactions are calculated according to a set of rules (atom types) and geometric relationships (distances, angles) between interacting atoms. For every active site atom or residue, a series of bits is then switched on or off as whether an interaction with the corresponding atom (atom-based IFP) or residue (residue-based IFP) is verified (Figure 14.3).

The interaction fingerprint concept (SIFt: structural interaction fingerprint) was pioneered by Biogen Idec. [24] for analyzing ligand docking poses to protein kinases, and showed several promising features: (i) enhancing the quality of pose prediction in docking experiments [24], (ii) clustering protein–ligand interactions for a panel of related inhibitors according to the diversity of their interactions with a target subfamily [25], and (iii) assisting target-biased library design [26]. IFPs were further developed by other groups in order to define the directionality of the
interactions (e.g., H-bonds donated by the ligand and by the active site are stored in distinct bits [14]), the strength of the interaction [27], or assign a bit to every active site atom instead of every active site residue [28]. In most applications, only frequently observed interaction types (hydrophobic, aromatic, hydrogen bond, ionic bond, metal complexation) are encoded in IFPs. Weaker and rare interactions (e.g., halogen bond, π–cation interaction) may however be described, provided that the interacting atoms and their topological relationships can be tabulated. For example, the CREDO database archives structural interaction fingerprints from 40,000 protein–ligand PDB structures using 23 different interaction types including nine ring–ring interaction geometries and three different atom–ring interaction types [29]. However, it should be stated that caution should be taken when trying to map rare interactions to IFPs. In many cases, the additional signal describing a particular interaction is spoiled by the noise that is concomitantly added upon registration of other weak interactions that would have been missed by standard fingerprints.

In virtual screening scenarios, we have been routinely postprocessing docking poses by IFP similarity to known actives in order to reconcile structure–activity relationships with plausible binding modes [30–34] and to identify novel ligands that would have been overlooked by classical energy scoring functions [35,36]. Remarkably, IFPs show a great scaffold hopping potential in selecting virtual hits sharing the same interaction pattern than a reference ligand, but with different chemotypes [37].

Figure 14.3 Converting a reference protein—ligand complex (Ref, left panel) into an interaction fingerprint (IFP) registering seven possible interactions with binding site residues. Docked poses (middle panel) are fingerprinted and the corresponding IFPs are compared with the reference IFP (Ref) using a standard Tanimoto coefficient. The best pose (pose 5, right panel) exhibits the highest IFP similarity to the reference.
14.4 True Target–Ligand Fingerprints

In the last decade, chemogenomic approaches to drug discovery have taken considerable importance, given the ever-increasing body of the available binding data on target–ligand complexes [38,39]. Instead of focusing on a particular target and/or chemical series, all available data (usually on a target family) are considered simultaneously in order to learn rules enabling the design of novel ligands for neighboring but poorly known targets. Chemogenomic approaches make an optimal use of true target–ligand fingerprints in which ligand and target descriptors (1D, 2D, 3D) are encoded into a single descriptor. Depending on which property (binary association, binding affinity, molecular interaction) has to be predicted, slightly different methods may be used.

14.4.1 Association Fingerprints

The association fingerprint (Figure 14.4) just describes a particular protein–ligand pair with no further information on which interactions are engaged and how strong they are.

Association of target–ligand fingerprints was originally designed to describe and predict G protein-coupled receptor (GPCR)–ligand pairing in chemogenomic applications [40–42]. In the fingerprint proposed by Weill et al. [42], ligand properties have been represented by standard descriptors (MACCS keys and SHED descriptors); protein cavities are encoded by a fixed-length bit string describing

![Figure 14.4](image-url)

**Figure 14.4** Protein-ligand association fingerprint: from a complex between a protein (tan ribbons) and a ligand (cyan sticks), ligand and target descriptors are concatenated into a single protein-ligand fingerprint.
pharmacophoric properties of cavity-lining amino acids. Several machine learning classification algorithms (SVM, random forest, naive Bayes) were trained on two sets of roughly 200,000 receptor–ligand fingerprints with a different definition of inactive decoys. Crossvalidated models show excellent precision (>90%) in distinguishing true from false pairs with a particular preference for random forest models. In most cases, predicting ligands for a given receptor was easier than predicting receptors for a given ligand. Protein and ligand descriptors need not be concatenated. Hence, separate kernels for ligands and targets may be developed as proposed by Vert and coworkers [41,43,44], the similarity in target–ligand space being expressed as the tensor product of pairwise ligand and target similarities. Interestingly, the most precise protein kernel (based on binding site 3D similarity, for example) is not necessarily the most reliable one. Bajorath and coworkers [45] suggested that simplified strategies for designing target–ligand SVM kernels should be used since varying the complexity of the target kernel does not influence the identification of ligands for virtually deorphanzied targets much. Hence, predicting protein–ligand association is clearly dominated by ligand neighborhood [45].

In the former applications, only target–ligand pairing was predicted. The same descriptors can however be used to predict the strength of the association, in other words, the binding affinity. In a pioneering work, Bock and Gough [40] used rather standard 2D topological and atomic descriptors for ligands, physicochemical properties of amino acid sequences for GPCRs, and concatenated feature vectors for both the receptor and the ligand in a single fingerprint. A support vector machine regression model was trained on 5319 receptor–ligand pairs from the PDSP Ki database [46] to predict the binding affinity of any ligand to any GPCR and to propose novel ligands for orphan receptors. Unfortunately, none of the predictions have ever been validated experimentally. Strömbergsson et al. applied the same concept to a wider target–ligand space considering either PDB [47] or DrugBank chemical spaces [48]. The best models achieved a remarkable accuracy of about 82% for validation and 75% for external prediction. In contradiction with previous studies [42,45], a pure ligand-based model led to a much lower accuracy, raising questions about the suitability of the chosen ligand descriptors. Hence, the utility of embedding target information into a chemogenomic classification vanishes rapidly when enough ligand information is already available for describing a target. In agreement with a previous work on three target classes (GPCRs, enzymes, ion channels) [41], our own experience with 7000 druggable protein–ligand complexes from the Protein DataBank suggests using chemogenomic models only if less than 30–50 ligands are available for each target [49].

Due to their novelty, there have been few successful applications of chemogenomic modeling. Weill et al. reported the use of concatenated protein and ligand fingerprints (PLFPs) to find nonpeptide oxytocin receptor ligands among a database of 330,000 commercially available compounds [50]. Interestingly, the chemogenomic method provided more validated hits than any other virtual screening method (2D and 3D similarity search) used in parallel. Using a sequence-based protein descriptor and structural ligand features, an SVM model was trained on
626 proteins and 10,000 active compounds to further discover, among a library of 85,000 drug-like compounds, novel active compounds for four pharmaceutical targets (GPR40, SIRT1, p38-alpha, and GSK-3β) [51]. A similar approach was reported to predict the binding affinity of non-nucleoside inhibitors to 14 mutants of the HIV-1 reverse transcriptase and therefore prioritize the most interesting leads for novel mutant sequences [52].

14.4.2 Interaction Pattern Fingerprints

In a typical interaction fingerprint (Section 14.3), a bit is defined for every active-site residue/atom. IFPs are therefore limited to analyze interactions with highly homologous active sites sharing a fixed number of cavity-lining residues/atoms. To overcome this drawback, coordinate-frame invariant fingerprints have been designed to register interaction patterns independently on the size of the binding site (Figure 14.5).

For example, cavity-independent fingerprints (APIF) [53] do not consider the absolute but the relative positions of pairs of protein–ligand interacting atoms and store information in a 294-bit fingerprint according to the interaction type and the distance between interacting pairs. Like standard IFPs, APIF scoring by comparison to known references was shown to outperform conventional energy-based scoring functions in docking-based virtual screening of compound libraries. Unfortunately, the obtained fingerprints are difficult to interpret since deconvoluting API into specific protein–ligand features or protein–ligand structural alignments is not possible.

**Figure 14.5** Ligand-annotated target fingerprints: from a complex of known 3D structure between a protein (tan ribbons) and a ligand (cyan sticks), interactions (violet lines) are encoded in a fingerprint with respect to their type and spatial relationships.
Along the same idea, Xie and Hwang [54] introduced the concept of protein–ligand interaction networks to score docking poses by a pure topological method. Nodes are represented by specific ligand and protein atom types whereas edges are created if these atoms are in direct interaction. To simplify the analysis, the final networks only contain motifs comprising two ligand atoms in interaction with three protein atoms. By generating networks for 6000 protein–ligand X-ray structures, networks are assigned frequencies. Upon docking a ligand, a topological score (MotifScore) rewards the frequently occurring networks whereas the rare networks are penalized. The scoring function was shown to perform the typical energy-based scoring functions [54]. Interestingly, some triads of protein atom nodes are enriched in ligand-binding sites, thereby providing a rationale for detecting protein cavities from atomic 3D coordinates [55].

PROLIX (Protein Ligand Interaction Explorer) is a database mining tool designed at Roche Laboratories, enabling the retrieval of protein–ligand X-ray structures, satisfying user-defined queries based on a list of required interaction types, interacting distances, and specific binding amino acids [56]. To achieve ultrafast searches, PROLIX uses a series of molecular fingerprints describing protein active sites (the binding environment) and residue-specific interactions (interaction types). A shell residue fingerprint (SRFP) first registers in a 40-position vector the frequency of interacting and noninteracting amino acids within a distance of 4.5 Å of the bound ligand. An interaction fingerprint then stores the frequency of nine interaction types (H-bond, ionic, cation–dipole, cation–pi, dipolar, halogen bond, donor–pi, pi–pi, and van der Waals) with each of the 20 possible amino acids. Finally, a graph-based representation of protein–ligand interactions enables the selection of the best matches in the Protein DataBank. Using a series of fingerprints of increasing complexities permits to rapidly eliminate impossible solutions, thereby restricting the search for the best matches to much fewer complexes but with a more accurate graph-based descriptor. The tool is not yet publicly available but enables very subtle queries for nonexperts in order to find protein–ligand complexes, verifying a very specific interaction pattern. Unfortunately, the tool does not provide a 3D alignment of protein–ligand structures along with a generic similarity score. This issue was recently addressed by Desaphy and Rognan, who designed novel interaction pattern descriptors (Figure 14.6) along with methods (Ishape, Grim) enabling protein–ligand structural alignment and interaction pattern comparisons [57].

Interactions (hydrophobic, aromatic, hydrogen bond, ionic bond, metal complexation) are detected on the fly and physically described by a pseudoatom centered between each pair of interacting atoms. Counting all possible triplets of interaction pseudoatoms within six distance ranges, and pruning the full integer vector to keep the most frequent triplets enables the definition of a simple (210 integers) and coordinate-frame invariant interaction pattern descriptor (TIFP) that can be applied to compare any pair of protein–ligand complexes. TIFP fingerprints have been calculated for 10 000 druggable protein–ligand complexes of the sc-PDB database [15] and have allowed the
systematic comparison of relationships between interaction pattern similarity and ligand or binding site similarity. We notably found that interaction pattern similarity strongly depends on binding site similarity [57]. In addition to the TIFP fingerprint, we developed two tools (Ishape, Grim) to align protein–ligand complexes from their intermolecular interactions and further quantify their similarity. Ishape is based on the overlap of interaction pseudoatoms using a smooth Gaussian function whereas Grim utilizes a standard clique detection algorithm to match interaction pattern graphs. Both tools are complementary and permit protein–ligand complex alignments capitalizing on either global or local interaction pattern similarities. The new fingerprint and companion alignment tools have been successfully used in three scenarios: (i) interaction-based alignment of protein–ligand complexes, (ii) postprocessing docking poses according to known interaction patterns, and (iii) virtual screening for bioisosteric scaffolds sharing similar interaction patterns (Figure 14.7).
14.5 Conclusions

This chapter illustrates the many ways of implicitly or explicitly encoding protein–ligand interactions into simple fingerprints for systematically registering and comparing interaction patterns among protein–ligand 3D structures. Since public data on binding affinities [58] and 3D structures [59] are increasing at an amazing pace, it is likely that more accurate fingerprints, describing a larger portion of the target–ligand landscape, will be developed in the future. Among the many applications in drug discovery that are likely to benefit from these developments is the high-throughput virtual screening for identifying compounds and or/targets verifying specific interaction patterns with their environment. Such virtual hits may lead to novel and patentable scaffolds that are truly bioisosteric to known references, and also to virtual ligand profiles permitting to decipher either main or secondary targets from phenotypic screens.
References


large crystal structure databases. *Journal of Chemical Information and Modeling*, 52, 1450–1461.


15
SkelGen

Nikolay P. Todorov

15.1
Introduction

SkelGen is a method and an associated software package that helps to generate ideas for novel protein ligands and scaffold hopping. It was initially developed at the Department of Pharmacology, University of Cambridge [1,2] and later marketed by De Novo Pharmaceuticals Ltd.\(^1\) [3]. The methodology was further developed and validated in collaboration with F. Hoffmann-La Roche AG [4]. Recently, the software has even been reimplemented and extended based on open-source technology by Molscape Research Ltd. with active site and scaffold-hopping constraints defined via a PyMOL\(^2\) graphical user interface plugin (Figure 15.1).

The following definition of the concept of scaffold hopping has been adopted for the purpose of this chapter. Scaffold hopping comprises methods and technologies concerned with finding isofunctional but structurally dissimilar molecular entities with specified activity [5]. Scaffold hopping is a special case of de novo design where some of the functional groups and fragments or constraints from the original ligand structure are retained in the derived (hopped) structures. Thus, in order to ensure isofunctionality, fragments from the ligand could be used or alternatively, atoms in the ligand or receptor could be used to define constraints that the designed ligands should satisfy.

The chapter summarizes the principles and methods behind the SkelGen software package, providing references to the primary literature for further details [2,4]. It then considers how de novo design could be tailored to scaffold hopping by defining ways to specify isofunctionality. The chapter is organized as follows. The methods for structure generation, scoring methods, optimization, and modes of application are described first, followed by an outline of experimental validation studies of the technology. The adoption of de novo design to scaffold hopping is presented with emphasis on using fixed fragments and pharmacophore site points

---

while consideration is also given to more the more complex issues of receptor flexibility, presence of tightly bound water molecules, and receptor specificity to the application of scaffold hopping.

15.2 Structure Generation and Optimization

Putative ligand structures are built within the constraints of the active site by linking predefined small fragments and are optimized by the method of simulated annealing.

15.2.1 Fragments and Fragment Sets

SkelGen has default fragment libraries derived from the analysis of ligands from the World Drug Index [6]. Frequently occurring fragments are retained. RECAP [7] is an interesting computational fragmentation methodology that splits input molecules according to known synthetic transformations (Figure 15.1). Terminal bonds are labeled according to the transform type (e.g., R3, R4, R11 in Figure 15.1); unique fragments are sorted by frequency of use and stored as building blocks. A table of compatible bond labels is defined and consulted during construction of new ligands (Figure 15.2).
Another useful fragment library is composed of acyclic and ring fragment sets (Figure 15.3). In order to balance the composition of the generated ligands, the acyclic and ring fragments are separated into different sets for selection. This separation is useful, since there are many ring systems compared with simple acyclic fragments; in addition, individual frequencies of use of different fragments are also specified.

The composition of the fragment library is under user control and is provided as input to the program. In addition to default sets, the user can also specify own fragments. This is particularly useful in the framework of scaffold hopping (Section 15.4).
15.2.2  
Structure Generation

An initial structure is built by linking together several fragments in a stepwise manner. This initial structure may contain atoms positioned in the forbidden regions of the receptor site or close intramolecular contacts. These are considered in the objective function and resolved during the optimization process.

Several structure transforms are applied to transition from structure to structure. The transforms include fragment additions, fragment removals, fragment replacements, flexible-bond rotations, rigid-body translations, and rigid-body rotations; random numbers are used to guide the selections.

Fragment connection rules, defined by the list of compatible bond labels (Figure 15.2), are followed during the construction process. For example, if a fragment is replaced by another fragment, it is checked that the labels of the bonds in the fragment are compatible to the labels of the bonds connecting the fragment to the rest of the structure [2]. The frequency of the transforms could also be modified and could affect the efficiency of the search. Before the transformed structure is scored, predefined rules for adjusting the conformation are applied and the structure is modified and refitted.

15.2.3  
Scoring and Optimization

SkelGen uses a penalty function to accommodate multiple constraint requirements imposed onto the ligand structure. The penalty score function becomes zero when all the constraints are satisfied and it is positive otherwise. Various constraint terms are included into the penalty function, the main ones being (i) inter and (ii) intramolecular interaction score, (iii) ligand efficiency, (iv) match of site point pharmacophore constraints, (v) requirements for chemical composition, and (vi) displacement of fragments from specified positions [2,4].

Once all the constraints are satisfied, the penalty score becomes zero and the corresponding solution is saved. Simulated annealing and parallel multiple copy simulated annealing [8] are used to optimize the ligand penalty score.

ScreenScore [9] is used as means to estimate ligand binding and ligand efficiency. When the requirement is to minimize ScreenScore (rather than achieving a score lower than a specified value), once an initial ScreenScore requirement is satisfied, the corresponding structure is recorded, ScreenScore is reset to a lower value, and the search is continued until the maximum number of iterations is reached.

15.2.4  
Ligand-Based Design

In the absence of 3D receptor structure, ligand-based design could be attempted when information is available about the ligands binding to the same active site.
The application of scaffold hopping in a ligand-based mode with SkelGen has also been demonstrated [10]. Ligand superposition or a single-bound conformation of a known active, a pseudoreceptor, is generated as a design envelope, within which structures are generated. Many of these structures have high similarity to known chemotypes. Scaffold hopping is readily achieved within this pseudoreceptor.

15.3 Validation Studies

Both retrospective validation and prospective experimental studies have been performed to establish whether SkelGen could reproduce known scaffolds and chemotype classes as well as suggest novel putative ligands.

15.3.1 Retrospective Validation Study: CDK2, COX2, ER, MMP-3

Retrospective validation study was performed by scientists at De Novo Pharmaceuticals Ltd. and F. Hoffmann-La Roche AG [4]. SkelGen was used to design ligand structures for four targets of pharmaceutical interest, namely, cyclin dependent kinase 2 (CDK2), estrogen receptor, cyclooxygenase-2 (COX-2) and stromelysin-1 (MMP3).

SkelGen was able to generate representatives of many inhibitor classes and the results demonstrated the importance of search constraint definition. All the examples used constraints derived from the analysis of known ligands, an important consideration in relation to the use of SkelGen for scaffold hopping.

15.3.2 Estrogen Receptor

In addition, experimental validation of SkelGen was later undertaken [11]. The objective was to demonstrate that SkelGen can design novel, synthetically tractable ligands with estrogen receptor α (ER-α) affinity. The starting points were seven ER-α crystal structures with known active compounds: 3ert, 1l2i, 1qkt, 1gwr, 1gwq, 1err, 1pcg.

According to the work plan, SkelGen was used to design ligands for each of the seven crystal structures. Two different strategies were used for each target and 1000 ligands were generated for each strategy, 140 000 structures in total.

Duplicates were removed and the top 25 structures ranked by ScreenScore were taken from each set of 1000. These structures were pooled and clustered, then checked for drug-likeness; synthetic feasibility and analysis were performed by a synthetic chemist. Thus, 35 high-scoring ligands were proposed for synthesis. Synthesis was performed at Peakdale Molecular. Compounds represented both the strategies and the diversity of SkelGen output. 17 compounds were synthesized.
and purified in the allotted time. Screening was carried out at Cerep using fluorescence polarization assay 813-ah, with the human recombinant ER-α and 17β-estradiol as reference compounds. IC₅₀s were determined for compounds with >40% inhibition at 10 mM. Five ligands, four of which are novel, showed affinity, with the best compound displaying an IC₅₀ of 340 nM (Table 15.1).

**Table 15.1 Designed compounds with ER activity.**

<table>
<thead>
<tr>
<th>Compound</th>
<th>IC₅₀ (μM)</th>
<th>MW</th>
<th>LE (kcal/mol/HA)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Compound 1" /></td>
<td>4.1</td>
<td>251</td>
<td>0.39</td>
</tr>
<tr>
<td><img src="image2.png" alt="Compound 2" /></td>
<td>24</td>
<td>242</td>
<td>0.37</td>
</tr>
<tr>
<td><img src="image3.png" alt="Compound 3" /></td>
<td>18</td>
<td>236</td>
<td>0.36</td>
</tr>
<tr>
<td><img src="image4.png" alt="Compound 4" /></td>
<td>25</td>
<td>268</td>
<td>0.32</td>
</tr>
<tr>
<td><img src="image5.png" alt="Compound 5" /></td>
<td>0.34</td>
<td>242</td>
<td>0.52</td>
</tr>
</tbody>
</table>

15.3.3 Histamine H3 Inverse Agonists

Roche and Sarmiento [12] used SkelGen in ligand-based mode in the design and synthesis of potent and selective nonimidazole inverse agonists for the histamine H3 receptor and to validate a new pharmacophore model. They arrived at four templates, which share the same pharmacophores but have different connectivity patterns and exemplified one of the patterns. The most promising compound was tested in a GTPγS functional assay and was found to be a potent inverse agonist with an EC₅₀ of 0.2 nM. Moreover, another compound also displayed a high affinity toward the H3 receptor with a Ki of 9.8 nM. Roche and Sarmiento demonstrated that the removal of the basic center common to both models leads to loss of activity, whereas the replacement of the second basic center by an acceptor retains the potency.
15.4 Scaffold Hopping Using Fixed Fragments

There are two main methods to tune SkelGen to perform scaffold hopping, instead of pure de novo design: (i) defining sets with fixed fragments in the fragment library and specified connection rules and (ii) using pharmacophore site points either from groups in the ligand or from complementary receptor groups in order to ensure isofunctionality with known scaffolds. This section describes method (i) while the next sections describe method (ii).

In SkelGen fragment input file, fragments can be divided into fragment sets and a specified number of fragments, between specified minimum and maximum value, can be selected from each set independently and incorporated into the ligand. This arrangement gives flexibility to control the composition of the fragments in the ligand.

Figure 15.4 shows an example of a fragment library definition in SkelGen with four fragment sets: one fragment from each set will be included into the ligand with sets 2–4 allowing choice between two different fragments (e.g., R1_00001.mol2 and R1_00002.mol2). For scaffold hopping, it is useful to be able to specify that certain fragments should be present in all structures considered during the search and all solutions, for example, R0_00001.mol2 in the example in Figure 15.4.

Fragment deviation (in terms of RMSD) could also be specified in the fragment library file and enforced with a penalty term in the scoring function. Thus, certain

```
4
-----------
1 1 1
./R0_00001.mol2 1.0 0 100 1 -1.0 1
-----------
2 1 1
./R1_00001.mol2 1.0 0 100 1 -1.0 1
./R1_00002.mol2 1.0 0 100 1 -1.0 1
-----------
2 1 1
./R2_00001.mol2 1.0 0 100 1 -1.0 1
./R2_00002.mol2 1.0 0 100 1 -1.0 1
-----------
2 1 1
./R3_00001.mol2 1.0 0 100 1 -1.0 1
./R3_00002.mol2 1.0 0 100 1 -1.0 1
```

Figure 15.4 Fragment sets.
fragments from a known ligand could be constrained to be present in all generated structures in positions close to those found originally.

Fragment connectivity rules (defined in input files, see Figure 15.2) could be used to specify exactly how fragments could be connected together. This option will limit the search over a more restricted structural space.

An example of the application of scaffold hopping using fixed fragments is shown in Figure 15.5 as applied to N-alpha-(2-naphthyl-sulphonyl-glycyl)-DL-p-amidinophenylalanyl-piperidine (NAPAP) ligand of Factor Xa (PDB code 1ppc, Figure 15.5a). Two fragments are selected from NAPAP, namely, benzamidine and naphtaline (Figure 15.5b). They are included as separate fragment sets in the SkelGen fragment library including RMSD constraints on their positions. One alternative scaffold generated by SkelGen is shown in Figure 15.5c.

As indicated by this example, scaffold hopping relates closely to the fragment linking concept from fragment-based drug design [13] (FBDD) and could be used for that purpose. Fragment growing is also related to scaffold hopping; here the scaffold to be replaced is at terminal position and has just one substituent.

It is also possible to apply a two-step procedure to de novo design [14]: (i) docking fragments and (ii) linking fragments. Here (ii) corresponds to scaffold hopping while (i) could be used to provide functionality if no known ligands are available.

This design approach to generate small fragments had been applied in the DNA gyrase ATP binding site using SkelGen. SkelGen was run to generate exhaustive number of structural possibilities, which were subsequently filtered for site complementarity and synthetic tractability. A number of known active fragments are found, but most of the species created are potentially novel and could be valuable for further elaboration and development into lead-like structures.

The deconstruction of lead-like inhibitors into putative fragments reveals the complexity of dealing with low-affinity leads, where fragments could have alternative binding modes compared with their positions in the lead structure, and gives rise to the concept of modular and emergent binding [15].

15.5 Scaffold Hopping Using Site Points

Site points, or pharmacophore constraints, are used to ensure crucial interactions between the ligand and the receptor. They could be associated with the ligand or the receptor and could be used to provide isofunctionality compared with known ligands. This could reduce the combinatorial space for structure generation.

Several types of site point constraints are usually used: (i) H-bonding atoms, (ii) hydrophobic atoms, (iii) atom element, (iv) aliphatic ring, (v) aromatic ring plane, (vi) inclusion/exclusion spheres. Site points can be grouped using OR and AND logical operations. This allows for efficient handling of large binding sites.

For example, hydrogen bond donor and acceptor points as well as hydrophobic points are defined in Figure 15.6 to define the strategies used in the ER study described in Section 15.3.2.
Figure 15.5  Fragment linking example. (a) Structure of NAPAP, (b) fragments derived from NAPAP (1 ppc), (c) new scaffold generated by SkelGen linking the fragments.
Site points could be derived from ligands with different hydrogen bonding patterns with the receptor. By using logical operations with the site points, it is possible to specify this extended isofunctionality pattern and derive a wider selection of scaffolds compared with requiring strict isofunctionality with a single ligand. It is also possible to have strategies with site points not matched by the original ligands and thus extend the size of scaffold-hopping space even further. SkelGen is not dependent on availability of ligands for the derivation of site points; alternative choices of site points are possible, covering the scaffold space between hopping and de novo design.

15.6 Further Considerations for Scaffold Hopping

Receptor flexibility, the presence of displaceable water molecules in the active site, and consideration of multiple proteins and aiming at ligand specificity or polypharmacology make scaffold hopping more complex to apply.

15.6.1 Receptor Flexibility

In many cases, the single static receptor conformation provides a limited description and several conformations or dynamic flexibility of the receptor should be considered. SkelGen has been adapted to use several static receptors and flexible side chains through its Reflex method [16,17].

Conformational ensemble could be derived using (i) experimental models: NMR, X-ray, (ii) homology models, (iii) molecular simulation, (iv) elastic net, (v) side
chain movements, (vi) rotamers, and (vii) backbone change. New scoring terms are added to the scoring function in order to take into account receptor–receptor interaction terms.

For the purposes of scaffold hopping, dynamic flexible site points associated with functional groups in the receptor could be defined as constraints for scaffold hopping. These site points follow the movements of the receptor and the ligand also have to adapt dynamically in order to satisfy the imposed constraints.

Scaffold hopping using fixed fragments is also more difficult to apply in the case of flexible receptor. It is possible to keep a fragment from a known ligand in its original position, but larger deviation from that position should be allowed and the binding energy requirements and dynamic site points should be relied on to make sure that the fragment adapts to fit the most suitable receptor conformation.

The approach has been applied to the design of ligands in the S1\textsuperscript{+} pocket of human collagenase (matrix metalloproteinase-1) MMP-1. The results of corresponding simulations for both rigid and flexible binding sites are compared in order to gauge the influence of receptor flexibility in drug discovery protocols and also reagent screening using a single-crystal structure and multiple NMR structures, both apo and holo forms [18].

The impact of receptor flexibility on the proposed solutions was also investigated, when applied to carboxypeptidase A, acetylcholinesterase, and the estrogen receptor (ER) [19]. The results showed that side chain movement in the active site generates solutions that were not accessible from the multiple static protein conformations available for these targets. Furthermore, an analysis of the structures generated in a flexible versus a static ER active site suggests that these additional solutions were not merely noise but contained many interesting chemotypes.

15.6.2 Water Molecules

The presence of displaceable water molecules in the binding site may influence scaffold-hopping strategies. Functional groups that displace water molecules could, in principle, be used as anchors for scaffold hopping in addition to functional groups in the ligands that form hydrogen bonds with the water molecules.

SkelGen has been applied to the X-ray crystal structure of bacterial neuraminidase in the presence of some selected water molecules [20–22]. The removal of all water molecules made generating potential ligands more difficult and the unsatisfied hydrogen bonding site points left by removing these water molecules were harder to satisfy by a ligand. It was easier to generate more diverse ligands if some of the water molecules were retained. Thus, water molecules make the binding site more accessible for hydrogen bonding and reduce the conformational constraints of a particular binding site. Energy calculations were conducted in order and were used to evaluate the consequences of the presence of tightly bound water molecules and to assess the ligand–protein and ligand–water–protein
interactions of different functional groups of the generated ligands to identify those that contribute favorably to the binding of ligands.

15.6.3
Receptor Specificity

Hopping of scaffolds binding to one receptor, but not binding to another, or binding to both receptors would be more restrictive compared with scaffold hopping to a single receptor. Structures could be designed first for a single receptor before consideration is given to the other receptor(s). This strategy reduces complexity, but could create multiple extraneous solutions that should be filtered at the later stage. Simultaneous ply-receptor scaffold hopping could take advantage of the use of dynamic site points and fixed fragments with large RMSD deviations, in a manner similar to that used for flexible receptor structures. Such a strategy could still be complicated since binding mode adjustment of the scaffolds to the different receptors should still be considered.

SkelGen was used to create ligands specific for the cyclin-dependent kinase (CDK) family [23]. In this work, we use novel computational approaches to design ligand candidates that are potentially inhibitory across the three CDK family members as well as more specific molecules toward specific members of the CDK family. According to a search of the World Drug Alerts, the highest scoring molecule that is specific for the all CDK family members shows similar chemical characteristics and functional groups to known active ligands.

15.7
Conclusion

Scaffold hopping is a promising method for the development of new lead compounds and SkelGen could be a useful tool for that purpose. The methods for structure generation, scoring methods, optimization, and modes of application as well as experimental validation of the proposed results were outlined. This chapter reviewed this method with particular emphasis on its application to scaffold hopping. SkelGen could be tailored by specifying isofunctionality with known ligand structures using (i) fixed fragments and (ii) pharmacophore site points. Consideration was also given to the more complex issues of receptor flexibility, the presence of tightly bound water molecules, and receptor specificity to the application of scaffold hopping.

Acknowledgments

I would like to thank Dr. Henriëtte Willems and Dr. Stuart Firth-Clark for the preparation of Figures 15.2 and Figure 15.6.
References

6 Thompson Scientific World Drug Index.

Part Three
Case Studies
Case Study 1: Scaffold Hopping for T-Type Calcium Channel and Glycine Transporter Type 1 Inhibitors
Leah C. Konkol, Timothy J. Senter, and Craig W. Lindsley

16.1 Introduction

Scaffold hopping, sometimes referred to as lead hopping, is a widely employed strategy in medicinal chemistry, which when used successfully allows rapid access to novel chemical structures, or chemotypes. These new structures may offer improved physiochemical properties and novel intellectual property (IP) space, while maintaining activity at the desired molecular target [1–9]. Scaffold hopping can be considered a subcategory of bioisosteric replacement, wherein scaffold hopping focuses on the replacement of the central core of a bioactive small molecule with a structurally dissimilar, yet topologically similar, chemotype [1–9]. The concept of scaffold hopping was introduced by Schneider et al. in 1999, and was heralded as a technique to identify bioisosteric replacements for the central cores of bioactive compounds, providing significantly different molecular backbones [10]. The concept of scaffold hopping as a named approach is young; however, medicinal chemists have been employing this strategy in lead optimization since the very beginning of drug discovery efforts, with many early marketed drugs being derived from other marketed drugs or natural products [3,4]. Recently, Sun et al. classified scaffold hopping strategies into four major categories: heterocycle replacements, ring opening or ring closure, peptide mimetics, and topology-based hopping, with the latter affording the greatest degree of structural novelty [2]. A multitude of software packages have been employed to aid the medicinal chemist in scaffold hopping efforts, utilizing topological, topographical, and surface-based (field-based) 2D and 3D descriptors [1–13]. The primary and patent literature is replete with examples of scaffold hopping jump starting drug discovery efforts and rapidly developing IP [1–10]. This chapter will focus on the application of scaffold hopping to develop novel cores as bioisosteric replacements for functionalized piperidines (a highly patented heterocyclic core), and through these exercises, access to novel IP space for both calcium T-type channel inhibitors [14] and glycine transporter type 1 (GlyT1) inhibitors [15–20], while maintaining the excellent target potency, drug metabolism and pharmacokinetic (DMPK)
16.2 T-Type Calcium Channel Inhibitors

T-type Ca\(^{2+}\) channels are voltage-gated calcium channels characterized by transient, or fast, inactivation of the channel. Molecular cloning studies have identified three members in the T-type calcium channel family: Ca\(_{\alpha}3.1\) (\(\alpha1G\)), Ca\(_{\alpha}3.2\) (\(\alpha1H\)), and Ca\(_{\alpha}3.3\) (\(\alpha1I\)) [27–38]. Inhibition of T-type Ca\(^{2+}\) channels hold tremendous therapeutic potential for the treatment of pain, epilepsy, sleep disorders, essential tremor, Parkinson’s disease, and other neurological disorders. However, target validation by small-molecule pharmacological inhibition of T-type Ca\(^{2+}\) channels has been hampered by either a lack of truly selective, central active small-molecule tool compounds or IP constraints [27–38]. The first generation of T-type Ca\(^{2+}\) channel inhibitors 1–3 were drugs designed for other targets (antiepileptic, antihypertensive, and antipsychotic), but ancillary pharmacology studies indicated weak inhibition of T-type Ca\(^{2+}\) channels (Figure 16.1); thus, these tools lacked the potency and selectivity necessary for target validation of T-type Ca\(^{2+}\) channel inhibition [27–38]. While these tools provided little in terms of our understanding of T-type Ca\(^{2+}\) channels, they did provide evidence that inhibition of T-type Ca\(^{2+}\) channels in man was well tolerated upon both acute and chronic dosing with 1–3. Against this backdrop, the NIH Molecular Libraries Roadmap launched the Molecular Libraries Screening Center Network (MLSCN) and the later evolved Molecular Libraries Production Center Network (MLPCN), which are nationwide consortiums of facilities that provide high-throughput small-molecule screening and medicinal chemistry for biological targets/pathways that lack small-molecule proof-of-concept tools.\(^1\)

1) For information on the MLPCN, see www.mli.nih.gov/mli/mlpcn.

Figure 16.1 First-generation T-type Ca\(^{2+}\) inhibitors derived from marketed antiepileptic 1, antihypertensive 2, and antipsychotic 3 drugs.

1, ethosuximide (antiepileptic) 2, mibefradil (antihypertensive) 3, pimozide (antipsychotic)
As evidenced by the known prior art with 1–3, T-type Ca\textsuperscript{2+} channels fit well within the MLSCN/MLPCN mandate, and therefore a T-type Ca\textsuperscript{2+} channel high-throughput screen (HTS) was performed in 2008 [38]. The screen was a kinetic assay utilizing HEK293 cells expressing the Cav3.2 channel against an ~111 000 member compound library measuring calcium fluorescence. After counterscreens of the 4246 hits, only 1 hit remained that was both active as a T-type Ca\textsuperscript{2+} channel antagonist and selective versus L- and N-type channels, a disubstituted 1,3,4-oxadiazole 4 (Cav3.2 IC\textsubscript{50} = 2.5 \mu M) [38]. While 4 was being confirmed, a number of reports appeared from the pharmaceutical industry describing potent and selective T-type Ca\textsuperscript{2+} channel inhibitors represented by 5–8 (Figure 16.2) [38,39]. Moreover, despite several displaying in vivo efficacy in various animal models, all had significant IP constraints associated with them. Based on this newly emerging prior art, the MLPCN charged our Center with developing a novel T-type Ca\textsuperscript{2+} channel inhibitor, with comparable potency, selectivity, and DMPK profile as 5–8, but free from IP constraints to empower the biomedical research community.

Due to the modular nature of the HTS hit 4, we employed parallel synthesis to quickly generate a diverse library of 56 analogs 9, wherein eight 2-amino-5-aryloxadiazoles were acylated with a diverse set of seven acid chlorides.
SAR within this library was flat, with all analogs possessing IC50 values of 1–5 µM, or inactive (IC50 values >10 µM) and in electrophysiology (EP) assays, all analogs 9 displayed a significant 10-fold leftward shift on potency; moreover, this series was far from the required potencies of 5–8. We then elected to hold the most optimal amide moiety constant and survey nonaromatic moieties in the 5-position of the oxadiazole. This library proved more effective providing our first novel submicromolar T-type Ca2+ channel inhibitor 10 (Ca3.2 IC50 = 0.8 µM); however, potency was again lost in the EP assay [38]. Clearly, this chemotype would not engender the necessary potency to become an MLPCN probe, but we began to notice structural similarities between our 10 and the T-type Ca2+ channel inhibitors 6 and 7 from Merck [21,22,38].

**Figure 16.3** Chemical optimization approach for VU/MLCPN HTS hit 4 and the most active analog 10.

(Figure 16.3) [38]. T-type Ca2+ channel inhibitors 6 and 7 from Merck originated from the des-fluoro analog 11 (Ca3.2 IC50 = 0.06 µM), which also possessed significant ancillary pharmacology, especially at ion channel targets [21,22]. Advanced analogs 6 and 7 maintained T-type Ca2+ channel activity, but the fluorine atoms inductively attenuated piperidine basicity from pKa 8.7 to 7.9, which improved ancillary ion channel pharmacology and engendered IP [21,22]. Examining molecular models of 10 and 11 indicated significant topological overlap. We then employed SurflexSim flexible alignment software [40] that further supported the topological similarity of 10 and 11, with the oxadiazole core mimicking the piperidine nucleus, but key hydrogen bond donors and acceptors were poorly aligned (Figure 16.4) [38]. Based on this observation, our strategy to access a T-type Ca2+ channel of comparable potency and DMPK profile to 6 and 7 resorted to a scaffold hopping exercise. Our goal was to access new IP space outside Merck, and to publish and not patent the composition, thereby making the probe freely available and free of IP constraints [38].

This insight led us to model other variations of 6 and 7 that would “hop” our series outside the Merck IP space [21,22]. One of the first series that seemed warranted centered on chiral 3-amino-substituted piperidines, as highlighted
generically by 12 and 13 (Figure 16.5) [38]. Several iterative libraries were prepared and tested, with the (R)-enantiomer 12 proving to be uniformly inactive.

Interestingly, the optimized 3,5-dichlorobenzamide of 6, 7, and 11 [21,22] provided weak antagonists within the (S)-enantiomeric core 13, but a 2-chloro-4-

![Figure 16.4](image1)

**Figure 16.4** Evolution of the Merck piperidine series of T-type Ca^{2+} channel antagonists. (a) Evolution of 6 and 7 from 11, and the structural similarity of 10 to 11. (b) SurflexSim flexible alignment of 10 (gray) and 11 (tan), showing similar topological similarity, but disparate H-bond donor–acceptor orientation.

![Figure 16.5](image2)

**Figure 16.5** Scaffold hopping into chiral 3-aminopiperidine cores 12 and 13. This series showed enantiospecific activity, with 14 (IC_{50} = 2.4 \mu M) as the most active analog in the series. A related [3.3.0] series 15 also displayed enantiospecific activity, but suffered from significant ancillary pharmacology.
fluorophenoxy amide provided 14, the most potent T-type Ca$^{2+}$ channel antagonist in this series [38]. However, the potency and ancillary pharmacology of 14 precluded its use as an MLPCN probe, though the enantiospecific activity was worthy of note. A related [3.3.0] series 15 also showed enantiospecific activity (as in the case of 14, only the $(S)$-enantiomer was active), but ancillary pharmacology was again an issue [38].

The enhanced potency with the [3.3.0] system in 15 led us to explore the contracted [3.1.0] bicyclic systems 16 (1R,5S,6S) and 17 (1R,5S,6R), as a perusal of patent databases indicated clean IP space (Figure 16.6) [38]. Moreover, we anticipated that the modest inductive effects of the $\pi$-character of the cyclopropyl ring of the [3.1.0] system would slightly diminish the $pK_a$ (8.1 for 16 and 17 versus 8.7 for 11) of the bicyclic nitrogen, and thus improve ancillary pharmacology.

Gratifyingly, enantiospecific inhibition was again observed, with 16 displaying robust T-type Ca$^{2+}$ channel inhibition ($IC_{50} = 150$ nM) not only in the calcium assay but also in the patch EP assays ($IC_{50} = 274$ nM) [38]. Importantly, the modest reduction in $pK_a$ translated to a clean ancillary pharmacology profile, with no significant activities (no inhibition $>50\% @ 10\mu M$) in a panel of 68 GPCRs, ion channels, and transporters, and the IP space was clear. It is important to note that scaffold hopping is not always a panacea. In the present case, we prepared a number of analogs around 16, but only the t-buty1 ethyl moiety, in combination with the 3,5-dichlorobenzamide, afforded potent T-type Ca$^{2+}$ channel inhibition (Scheme 16.1). Of note, the synthesis was rapid, as the [3.1.0] core 18 was commercially available [38].

![Figure 16.6](image)

**Figure 16.6** Scaffold hopping into [3.1.0] cores as a bioisostere for the piperidine scaffold. (a) The (1R,5S,6S) isomer 16 displayed potent T-type Ca$^{2+}$ channel inhibition ($IC_{50} = 150$ nM), and SurflexSim flexible alignment of 16 (gray) with 11 (tan) suggests good topological overlap. (b) The (1R,5S,6R) isomer 17 displayed weak T-type Ca$^{2+}$ channel inhibition ($IC_{50} = 5\mu M$), and SurflexSim flexible alignment of 17 (gray) with 11 (tan) suggests poor topological overlap.

![Scheme 16.1](image)

**Scheme 16.1** Synthetic route to access libraries of analogs of 16, starting from commercial 18.
The next hurdle toward declaring 16 an MLPCN probe was to compare its DMPK profile with that of 6 and 7. Here as well, scaffold hopping established 16 immediately as a favorable lead compound. In plasma protein binding studies, 16 possessed an exceptional free fraction in both rat ($f_u = 9.1\%$) and human ($f_u = 3.3\%$) and a clean CYP450 (>30 μM versus 3A4, 2C9, 1A2). A good in vitro: in vivo correlation was noted, with an in vivo plasma clearance value [CL$p = 56 \text{ ml/(min kg)}$] that correlated with the predicted hepatic clearance value [CL$_{HEP} = 43 \text{ ml/(min/kg)}$] obtained from the rat liver microsomes, yet providing a 7 h half-life. Importantly, 16 was highly brain penetrant with a BrainAUC/PlasmaAUC of 7.4. Overall, the in vitro and in vivo profile of 16 was virtually identical to Merck’s 6 and 7, which showed modest rat PK parameters, but excellent higher species and predicted human PK parameters. In a well-established pharmacodynamics model of Parkinson’s disease, haloperidol-induced catalepsy, oral administration of 16 was identical in efficacy across a full dose range (0.3–56.6 mg/kg p.o.) to that of Merck’s 6 and 7. In addition, a satellite pharmacokinetic study showed tight, linear increases in exposure with dose of 16. Based on these data, 16 was declared an MLPCN probe, coined ML218, and was highly comparable to Merck’s 6 and 7 [21,22] in terms of potency, selectivity, PK, and PD; however, it was in clear IP space and thus freely available to the biomedical research community [38]. For this project, scaffold hopping was an essential tool for the medicinal chemist to overcome a poorly tractable HTS lead. Ligands were rapidly accessed with exceptional properties in new chemical space in less than 2 months and with limited human resources.

16.4 Scaffold Hopping to Access Novel Glycine Transporter Type 1 (GlyT1) Inhibitors

Recent publications suggest that a great deal of pharmaceutical industry and academic interest exists in the identification and development of selective GlyT1 inhibitors for the treatment of the cognitive and negative symptoms of schizophrenia by modulating the NMDA receptor (the so-called NMDA receptor hypofunction hypothesis of schizophrenia) [15–26,41–50]. These efforts have led to the development of numerous and structurally diverse GlyT1 inhibitors that are classified as sarcosine and nonsarcosine based; moreover, many have been in the clinic, and efficacy has been demonstrated [15–26,41–50]. As with the T-type calcium channel antagonists, the IP space is highly congested, and once again, Merck published and patented a number of variants with a central 4,4-disubstituted piperidine core – an ideal scaffold form to scaffold hop [38,42–50].

Merck’s effort began with an HTS effort that identified 21 [23], which evolved into related congeners 22–24 [24–26] and ultimately the clinical candidate 25, MK-2637 (Figure 16.7) [51]. Here, we employed the same scaffold hopping strategies described previously, and were rapidly able to develop four patented series of GlyT1 inhibitors, represented by 26–28 (Figure 16.8) [15–20]. In this instance, the piperidine core was effectively replaced by homologated azetidines/
piperidines 26, spirocyclic congeners 27, the [3.1.0] core 28, and the [3.3.0] core 29 [15–20]. All four cases withstood patent examination, and the US patents are beginning to issue [15,16]. Similar to the T-type calcium antagonists described earlier, these series displayed enantiospecific activity [15–20] and provided rapid access to potent GlyT1 inhibitors with comparable potency (IC50 values from 750 pM to 300 nM), selectivity, DMPK profiles, and pharmacodynamics profiles to those developed by Merck [23–26,51]. Clear themes emerged as bioisosteric replacements for the piperidine core, and solid IP resulted once again.

Figure 16.7: Merck’s 4,4-disubstituted piperidine series of GlyT1 inhibitors 21–24, leading ultimately to the clinical development candidate 25, MK-2637.

Figure 16.8: Vanderbilt GlyT1 inhibitors 26–29 derived from a scaffold hopping exercise of the 4,4-disubstituted piperidine-based GlyT1 inhibitors 21–24 from Merck. The effort afforded rapid access to potent and selective ligands with excellent DMPK and in vivo profiles.
16.5 Conclusions

In this chapter, we hope that we have demonstrated that scaffold hopping is an effective strategy in medicinal chemistry to rapidly access novel chemical structures, or chemotypes, that may offer improved physiochemical properties and novel IP space, while maintaining activity at the desired molecular target. Moreover, scaffold hopping can reduce the time required to develop a robust \textit{in vivo} tool compound or clinical candidate by years. Here, we showed two case studies where a piperidine core was effectively replaced by a variety of novel cores. These changes engendered enantiospecific activity, excellent potency and selectivity, good DMPK profiles, and \textit{in vivo} activity comparable to the patented leads from which they were derived. Without question, scaffold hopping is a powerful component of the medicinal chemist’s arsenal.

References


Case Study 2: Bioisosteric Replacements for the Neurokinin 1 Receptor (NK1R)

Francesca Perruccio

17.1 Introduction

A bioisostere is a compound resulting from the exchange of an atom or a group of atoms with another, broadly similar, atom or group of atoms. The objective of a bioisosteric replacement is to create a new compound with similar biological properties to the parent compound. This approach has been proven as a valid technique in drug discovery and in the last few years, a significant number of software have released or improved their own methodology to perform bioisosteric replacements in small molecules [1]. Features such as excluded volumes, shape of the binding site (when the putative binding site in the protein structure is known and available), geometry, and minimization of the final virtual compounds are now accessible in many computational tools. Moreover, most of these software, aiming for scaffold hopping, also provide up-to-date fragment collections (e.g., based on the most common fragments found in vendor compounds and commercial drugs) and the possibility to replenish them with in-house small molecules (e.g., by defragmentation of the company chemical compound material) [2]. When reporting a study of a successful case of bioisosteric replacement, it is often tricky to pick the right target to investigate and then discuss. For this chapter, neurokinin 1 receptor (NK₁R) has been chosen for this purpose: neurokinin 1 receptor is a well-known target in the already published data that can be used as a useful source of information in this retrospective study on bioisosteric replacement [3–5]. Therefore, NK₁ receptor appears to be an interesting target: more than one successful bioisosteric replacement has been reported in literature and this information can be used in the discussion of the results of this chapter.

17.2 Neurokinin 1 (NK₁) Therapeutic Areas

In general, drugs have a single therapeutic effect, but less frequently, they may exert several. Neurokinin 1 receptor antagonists are promising drugs, exerting...
many therapeutic effects with a dose-dependent action that depending on the concentration has more than one positive effects.

Neurokinin 1 is the receptor that mainly mediates the biological actions of the neuropeptide substance P (SP) [6]. The neuropeptide substance P is an undecapeptide that shows a widespread distribution in both the central and peripheral nervous systems. Many studies have also revealed a widespread distribution of the NK1 receptor in the mammalian central nervous system and also in peripheral tissues, vascular endothelial cells, muscle cells, gastrointestinal and genitourinary tracts, lung, thyroid gland, and immune cells [7].

After binding to the NK1 receptors, the neuropeptide SP can regulate many biological functions in the central nervous system from stress, depression, anxiety, nausea, migraine, alcohol addiction, and generic emotional behavior to progressive loss of structure or function of neurons, including death of neurons. The neuropeptide SP has been also implicated in pain, inflammation, chemical-driven liver damage, and virus proliferation. The neuropeptide SP also plays an important role in cancer (e.g., tumor cell proliferation, formation of new blood vessels, and migration of tumor cells for invasion and metastasis) [8–16]. However, NK1 receptor antagonists specifically inhibit the above biological functions mediated by the neuropeptide SP. Therefore, these antagonists can inhibit biological functions in the central nervous system and act as an anxyolitic, antidepressant, antiemetic, antimigraine, antialcohol addiction, or as a neuroprotector. They can also play a role in analgesic and antiinflammatory functions, preventing liver damage processes and in antivirus proliferation [9–21]. Regarding cancer, NK1 receptor antagonists have an antitumor action (inducing tumor cell death by apoptosis), and induce antiangiogenesis and inhibit the migration of tumor cells [22–40].

17.3 The Neurokinin 1 Receptor (NK1R) and Its Mechanism

The neurokinin 1 receptor (also known as tachykinin receptor 1 (TACR1) or substance P receptor (SPR)) is a G protein-coupled receptor found in the central and peripheral nervous systems. The endogenous ligand for this receptor is the neuropeptide SP, although it has some affinity for other neuropeptides. The neuropeptide substance P belongs to one of the largest families of neuropeptides (tachykinin peptides, so called due to their ability to rapidly induce contraction of the gut tissue [41]), found from amphibians to mammals. This neuropeptide family is characterized by the same hydrophobic C-terminal region with the amino acid sequence Phe-X-Gly-Leu-Met-NH2, where X is either an aromatic or a beta-branched aliphatic amino acid residue. The N-terminal region, instead, varies between different members of the tachykinin family [42–44]. The neuropeptide SP is the most interesting and therefore the most studied member of the tachykinin family. It is an undecapeptide with the amino acid sequence Arg-Pro-Lys-Pro-Gln-Gln-Phe-Phe-Gly-Leu-Met-NH2 [42]. The neuropeptide substance P can bind to all three of the tachykinin receptors (NK1, NK2, NK3), but it binds most strongly to the
NK₁ receptor [43]. The neurokinin 1 receptor [45] consists of 407 amino acid residues, and it has a molecular weight of 58,000 [42,46]. The neurokinin 1 receptor, as well as the other tachykinin receptors, is made of seven hydrophobic transmembrane (TM) domains with three extracellular and three intracellular loops, an amino-terminus, and a cytoplasmic carboxy-terminus (see Figure 17.1). Amino acid residues that are responsible for the binding of the neuropeptide substance and its antagonists are present in the extracellular loops and transmembrane regions of the neurokinin 1 receptor.

17.4 Neurokinin 1 Antagonists

Neurokinin 1 receptor antagonists specifically inhibit a variety of biological functions mediated by the neuropeptide SP, spanning from mood disorders, stress, anxiety, vomiting, and pain to tumor cells proliferation and mobilization. The inhibition effect triggered by NK₁ receptor antagonists has been exploited in the research and development of a class of medications that possesses unique antidepressant, anxiolytic, and antiemetic properties. Neurokinin 1 receptor antagonists can be classified in two classes: peptide antagonists and nonpeptide antagonists. Over the past few decades, the research into and development of antagonists of the NK₁ receptor have offered an important opportunity for the development of novel therapeutic agents. Several peptide antagonists derived from SP have been developed. The usefulness of these antagonists is limited in in vivo studies for the following reasons: their inability to gain access to the central nervous system through the blood–brain barrier, their neurotoxicity after administration in the
central nervous system, their mast cell degranulation activity, their affinity is several orders of magnitude lower than that of natural agonists, their partial residual agonist activity, their metabolic instability, and their poor potency and ability to discriminate between tachykinin receptors. In the last two decades, nonpeptide based NK₁ receptor antagonists have been subjects of interest in many pharmaceutical companies, as reported in the literature [47–50]. An overview of the most well-known NK₁ nonpeptide antagonists in the pharmaceutical scenario is reported in Table 17.1. In 1991, three research groups were involved in identifying NK₁ receptor antagonists by screening of compounds collections. Eastman Kodak and Sterling Winthrop discovered steroid series of tachykinin NK₁ receptor antagonists that yielded some compounds but lacked sufficient affinity for the NK₁ receptor then and

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Name</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Molecule 1" /></td>
<td>CP-96345</td>
<td>Pfizer</td>
</tr>
<tr>
<td><img src="image2.png" alt="Molecule 2" /></td>
<td>CP-99994</td>
<td>Pfizer</td>
</tr>
<tr>
<td><img src="image3.png" alt="Molecule 3" /></td>
<td>L-733060</td>
<td>Merck</td>
</tr>
<tr>
<td><img src="image4.png" alt="Molecule 4" /></td>
<td>MK-869</td>
<td>Merck</td>
</tr>
<tr>
<td><img src="image5.png" alt="Molecule 5" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
no improvement was achieved in terms of activity in consequent studies [50,51]. The steroid series for the NK₁ receptor also proved to be toxic. Rhône-Poulenc discovered a series with high affinity for the neurokinin 1 receptor in rats and mice, but not in humans. An improvement of the selectivity for the human NK₁ receptor resulted in the development of a compound with good activity in vivo and in models of pain and was developed up to phase II for the treatment of migraines but was then terminated [50,51]. In the same years, Pfizer discovered a benzylamino quinuclidine scaffold (CP-96345, Table 17.1) and its analog (CP-99994, Table 17.1). CP-99994 entered phase II clinical trials; these were later discontinued because of poor bioavailability. A large number of nonpeptide neurokinin 1 receptor antagonists have followed these first ones. L-733060 (Table 17.1) and derivatives are examples of a compound that was developed from CP-99994. Further studies conducted by Merck on this molecular scaffold yielded the compound MK-869, which showed high affinity for the NK₁ receptor and high oral activity (Table 17.1). MK-869 is also called
aprepitant, and was studied in pain, migraines, emesis, and psychiatric disorders. These studies led to the FDA-approved drug Emend for chemotherapy-induced nausea and vomiting, which is available for oral use [50]. A watersoluble phosphoryl prodrug for intravenous use, called Fosaprepitant, is also available and is marketed as Ivenend [52]. Aprepitant was also believed to be effective in the treatment of depression. It entered phase III trials before the development for this indication was discontinued [50]. Many different compounds have been described by various pharmaceutical companies besides the compounds that led to the discovery of Aprepitant. GR-205171 (Table 17.1) was developed by Glaxo and was also based on CP-99994. GR-205171 was developed up to phase II clinical trials for the treatment of postoperative nausea and vomiting, migraine, and motion sickness. It showed good results in emesis, but its development was discontinued [50]. LY-303870, or lanepitant (Table 17.1), is an N-acetylated reduced amide of L-tryptophan (Table 17.1) that was discovered by Eli Lilly. It underwent phase II clinical trials for the treatment of osteoarthritis pain but showed no significant effects. Eli Lilly carried on SAR work on its structure and developed compounds that did not enter clinical trials [50]. By a general hypothesis on peptidergic G protein-coupled receptors binding site, Takeda discovered a series of N-benzyl carboxyamides in 1995. One of those compounds, TAK-637 (Table 17.1), underwent phase II clinical trials for urinary incontinence, depression, and irritable bowel syndrome, but the development was discontinued. More recent studies have focused on nonpeptide antagonist of the tachykinin NK1 family to address the major problem of chemotherapy-induced emesis in cancer treatment. A new compound, T-2328 (Table 17.1), is administered intravenously, and treats both acute and delayed emesis. It is proposed to exert its very potent antiemetic effect through acting on brain NK1 receptors. In 2009, it was found that the NK1R gene is transcribed in four mRNA variants, which at the same time are translated into four NK1R isoforms that have different affinities to substance P. A study on these genetic variants of NK1R showed that these polymorphisms of the NK1R gene are associated with the development of alcohol dependence [53].

17.5 NK1 Receptor: Target Active Site and Binding Mode

There is more than one ligand-binding domain on the NK1 receptor for the nonpeptide antagonists, and these binding domains can be found in various places. The main ligand-binding site is in the hydrophobic core between the loops and the outer segments of transmembrane domains 3–7 (TM3–TM7) [52]. Several residues, such as Gln165 (TM4), His197 (TM5), His265 (TM6), and Tyr287 (TM7) are involved in the binding of many nonpeptide antagonists of the NK1 receptors [46,52]. It has been stated that Ala-replacement of His197 decreases the binding affinity of CP-96345 for the NK1 receptor. Further site-directed mutagenesis
experiments have shown that replacing Val116 (TM3) and Ile290 (TM7) decreases the binding affinity of certain antagonists, but this effect is rather an indirect influence of the overall conformation of the antagonist-binding site. The residue Gln165 (TM4) has also proven to be meaningful for the binding of several nonpeptide antagonists, possibly through the formation of a hydrogen bond [51,55]. Phe268 and Tyr287 have been proposed as possible contact points for both agonist and antagonist binding domains [46]. The significance of His265 has been confirmed in the binding of antagonists to NK1 receptor. His265 has also been reported to interact favorably with ligands, but at the same time, it has been demonstrated that Ala-replacement of His265 does not affect their binding affinity [49]. Some other residues that are thought to be involved in the binding of nonpeptide antagonists to NK1 receptor are Ser169, Glu193, Lys194, Phe264, Phe267, Pro271, and Tyr272. Each structural class of nonpeptide NK1 receptor antagonists appears to interact with a specific set of residues within the common binding pocket. [46,52]. A schematic diagram of proposed interactions (based on site-directed mutagenesis and published data) of CP-99994 with the human NK1 receptor is reported below (Figure 17.2). The benzyl amino group of CP-99994 is proposed to hydrogen bond with Gln-165. His-197 forms an amino aromatic interaction with the benzyl ring in the C-3 position. His-197 forms an amino aromatic interaction with the phenyl moiety in the C-2 position of CP-99994.

In 2010, the NMR structure of the neuropeptide SP was released. The corresponding publication provided insights on the possible interactions between the undecapeptide and the neurokinin 1 receptor based on a homology model of the latter. No crystal structure of the target enzyme has been solved yet.

![Figure 17.2](image_url) Proposed binding mode of CP-99994.
17.6 Bioisosteric Replacements in NK₁ Receptor Antagonist

The case study reported in this chapter is based on published data in literature on a bioisosteric replacement in the tryptophan-derived NK₁ receptor antagonist [56]. This case study represents an example of a nonclassical bioisosteric replacement where a noncyclic functionality of the molecule under investigation can be substituted by a cyclic group resulting in retention of biological activity. The aim of discussing this particular bioisosteric replacement for this given enzyme is to highlight the useful bioisosteric replacement, the importance of the geometry of the ligand under investigation, and how a commercially available computational tool would perform in this task. Given the fact that the publication used as the starting point for this case study is already dated a few years back, is very likely that the “right” answer will be contained in the databases queried by the software, giving us the possibility to compare the results presented in the publication with the outcome of the computed calculations. The case study discussed here is based on heterocyclic bioisosteres of the ester linkage in tryptophan-derived NK₁ antagonists. The discovery of (S)-tryptophan benzyl esters such as L-732138 as a class of potent and selective NK₁ antagonists has been published by Merck (see Figure 17.3). In this molecule, the ester linkage was known to contribute significantly to receptor binding, presumably via hydrogen bonding. However, the ester linkage also represented a possible weakness to be subject in vivo to the action of esterases. The distribution of esterases in the body is ubiquitous and several types can be found in the blood, liver, and other organs and tissues. In many instances, the prevalence of these esterases causes these molecules to be highly labile in vivo. With the aim to improve the in vivo stability of this class, further studies were carried out by Merck to introduce more robust linking groups into the tryptophan backbone of the lead ester series.

Studies of the available X-ray crystallographic data, and molecular modeling of the Pfizer NK₁ antagonists, have led to postulation of the existence of an intramolecular π–π interaction between the pendant aromatic moieties [57–61]. In selecting the heterocyclic templates for investigation, a primary objective was that a π–π interaction between the indole and bis(trifluoromethyl)phenyl rings should be

![Figure 17.3](image-url)

Figure 17.3 L-732138, IC₅₀: 2.5 nM.
potentially accessible. As a starting point, 2,5-dioxopiperazine was investigated as a suitable choice for preliminary studies, since all the connectivity of the backbone of the original ester lead seemed to be retained, including the important S-stereochemistry of the achiral center. This ring system also offered potential sites for further exploration by the subsequent attachment of substituents. 2,5-dioxopiperazine resulted to have a very weak receptor affinity (compound 2 in Table 17.2), but 2-oxopiperazine (compound 5 in Table 17.2) exhibited modest activity. Molecular modeling was performed to investigate the conformation of the

<table>
<thead>
<tr>
<th>Index</th>
<th>R1</th>
<th>R2</th>
<th>Linkage</th>
<th>Human NK₁ receptor binding affinity (a) (nM)</th>
<th>Comments</th>
</tr>
</thead>
</table>
| 1     | ![Image](image1.png) | ![Image](image2.png) | O=O-R₂-O⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻ putchar(`\n`) \(\text{Continued}\)
piperazine system as a replacement for the ester link. It is apparent that the central
piperazine rings of these two molecules occupy planes that are almost orthogonal.
These observations suggested the possibility of deleterious interactions between
the receptor and part of the piperazine ring system as a cause of the differences
in binding affinity of these molecules. Examination of the conformations accessible
to the substituents on these templates led to exploration of the corresponding
five-membered heterocyclic rings. This study culminated in the identification
of oxazolidinedione as a suitable ester mimic in terms of the retention of
NK₁ binding affinity. This approached gave the following results (see scheme
below in Table 17.2).

<table>
<thead>
<tr>
<th>Index</th>
<th>R1</th>
<th>R2</th>
<th>Linkage</th>
<th>Human NK₁ receptor binding affinity a (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td>317 ± 62</td>
</tr>
<tr>
<td>8</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td>153 ± 5</td>
</tr>
<tr>
<td>9</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td>63 ± 5</td>
</tr>
<tr>
<td>10</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td>22 ± 7</td>
</tr>
<tr>
<td>11</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td>11 ± 2</td>
</tr>
<tr>
<td>12</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td>31 ± 12</td>
</tr>
</tbody>
</table>
An alternative strategy to look for bioisosteres of the ester linkage in tryptophan-derived NK₁ antagonists was to alter the ester moiety to an amide or ether. This led to a substantial loss in binding affinity, but conversion to a ketone provided compounds with affinity comparable to the equivalent esters [62]. The findings of this study are summarized in the scheme below (Figure 17.4).

The resulting structure was optimized to improve solubility and bioavailability, leading to L-737488 (Figure 17.5) that bears a quinuclidine basic moiety. L-737488 showed high in vivo activity (ID₅₀ = 1.8 mg/Kg p.o. in inhibiting plasma protein extravasation in guinea-pig), weak affinity for Ca²⁺ channels, and good solubility in water [62]. No further development was reported for the series.

These studies identified interesting and equipotent bioisosteres regardless of the difficulties to attain optimal binding geometry. In the specific case of a cyclic group as possible replacement of an ester linkage, the ring systems investigated could
represent useful bioisosteric replacements for an ester moiety within another series of compounds. Moreover, unlike the ester moiety, these bioisosteres would not be subject in vivo to the action of esterases.

17.7 Bioisosteric Replacements in NK₁ Receptor Antagonist: A Retrospective Study

In the past years, there have been several attempts to examine sets of known active compounds to empirically identify bioisosteric fragments. This approach, however, can only suggest bioisosteric fragment pairs that are already known and it also often identifies many incidental rather than meaningful bioisosteric replacements. An alternative approach has been to generate an algorithm that would predict whether the two fragments are bioisosteres [63–67], and several groups have developed methods in this area. In the past few years, the quest for bioisosteric replacements has been of great interest of many software companies. The modern trend when looking at computational tools to perform bioisosteric replacements is the description of the moiety investigated by chemoinformatic techniques that can characterize groups by a range of calculated descriptors and identify bioisosteric pairs based on similarity of properties [68]. In this section, the computational tool SparkV10 has been applied to L-732138 with the aim to reproduce the previously discussed findings on the bioisosteric replacement of the ester linkage in tryptophan-derived NK₁ receptor antagonists (Figure 17.6). The purpose of this approach is to have an idea of which outcome we could obtain using this software and to query its fragments database. The reason behind the choice of this particular software is simply the fact that it is commercially available: the calculations have been run on the generic database that is provided in the software package making this study not dependent on in-house scripts and/or in-house collection of chemical fragments. SparkV10 (also known as the previous name of FieldStere) is a computational methodology developed and distributed by the software company Cresset [69]. SparkV10 uses Cresset’s field technology to find biologically equivalent

Figure 17.6 Conformation of the template molecule, L-732138.
replacements for a selected moiety in the molecule under investigation, enabling the user to find new structures in new chemical space. Cresset calculates four molecular fields to represent the binding properties of a ligand. Probes are located around the molecule under investigation to identify the following interactions: positive electrostatic (reported in red), negative electrostatic (reported in blue), van der Waals attractive that is, “steric” (reported in yellow), and hydrophobic (reported in orange) (Figure 17.7). The combination of the greatest interaction around the ligand gives the calculated pharmacophore. As in many methods based on 3D pharmacophoric features, the search for a suitable conformation of the template molecule is a key step of the scaffold hopping approach. The conformation of the molecule under investigation, L-732138, was investigated using standard MacroModel conformational search application in Maestro [70]. Among all the conformations accessible to the molecule, the majority of conformations presented the existence of an intramolecular π–π interaction between the pendant aromatic moieties with a number of higher energy variants when this interaction is not present anymore. A search in ConQuest [71] to retrieve information from the Cambridge Structural Database (CSD) gave further insights on the possible bioactive conformation of the molecule under investigation. The template molecule was therefore imported in SparkV10 in the conformation reported in Figure 17.5 to perform the bioisosteric search.

When imported in SparkV10, the molecular fields are calculated and the location of greatest interaction points between the template molecule and
charged and neutral probes is displayed (positive electrostatic interactions are reported in red, negative electrostatic interaction are reported in blue, van der Waals attractive, that is, “steric” interactions are reported in yellow, and hydrophobic interactions are reported in orange). The combination of these points describes in a rather sophisticated way the pharmacophoric features of the molecule under study (Figure 17.7).

SparkV10 identifies strong acceptor features of both the carbonyl groups of the ester linkage and the amide substituents. The same feature is not present, instead, for the oxygen atom of the ester linkage. Regarding the donor features, these appear to be more apparent for the nitrogen within the tryptophan ring rather than the nitrogen belonging to the amide group. The hydrophobic regions highlighted by the software are as expected in pendant aromatic moieties. SparkV10 allows the user to manually select the region of the template molecule to replace: for this retrospective study, we have selected the ester linkage of the ligand. It is important to anticipate that in this specific case, the software will suggest replacements and will still be able to have an appropriate attachment point for the acetamide side chain. The query has been carried out on commercial fragments databases provided within the software (very common 5533 fragments, common 18759 fragments, less common 24151 fragments, rare 44977 fragments, and very rare 48775 fragments) with the default option to retain the top 500 resulting molecules (ranked by the index score, the combination of the field score and the shape score). The calculations are not time demanding. The outcome of this search is a collection of possible ideas to replace the ester linkage of the molecule under investigation: when postprocessing these results, it is important to understand the diversity within the suggested virtual compounds and the corresponding chemical feasibility. It is not to be expected that a chemist will synthesize 500 singletons as possible bioisosteric replacements; instead, the selection and further design of which molecules to follow up is a key step in the quest for a bioisosteric replacement. In the case discussed in this section, the run performed with SparkV10 has resulted in a collection of 500 possible bioisosteric replacements: a simple cluster algorithm can already give us an idea of the recurrence of certain scaffolds. Within the top 50 scored molecules, the scaffolds suggested by the software SparkV10 to replace the ester linkage can be grouped as open chain linkers and heterocyclic linkers. The most recurrent open chain linkers are the ketone group and the amide. Both of these scaffolds seem to be able to retain the geometry of the starting molecule, and both of these scaffolds also retain the strong acceptor features of the carbonyl groups identified by SparkV10 in the ester linkage. However, from the published data, it is known that the ketone linkage replacement can give an active compound, whereas the amide is inactive. There might be a possibility that the amide group could also present an unfavorable interaction with the target enzyme. Within the top 50 scored molecules, the most recurrent heterocyclic scaffold also suggested by the software SparkV10, the five-member ring with diverse substitution patterns is not only able to retain
the geometry but also to fulfill the pharmacophoric features (strong acceptor features of the carbonyl groups of the ester linkage) highlighted by the software. It is interesting to notice that no six-member rings appear to be a suitable scaffold to replace the ester linkage. Overall, in the top 500 virtual compounds suggested by the software as possible bioisosteric replacements of the ester linkage of the molecule under investigation, we could easily retrieve three interesting starting points for possible chemical synthesis. This selection was made by clustering the proposed scaffolds, and addressing synthesis feasibility. By making use of the published data, we could link these virtual molecules to their in vitro activity reported in literature, and have an idea of the possible outcome of this run of calculations. The results of these calculations are reported in Table 17.3.

Compound 3 has been not reported in literature (with the acetamide side chain); however, the closest molecules that we could compare it with are compound 11 and compound 12 in Table 17.2, with human NK₁ receptor binding affinity of 11 nM and 31 nM, respectively. When compared with the template molecule L-732138, a bioisosteric replacement such as the ketone did retain a comparable level of biological activity, whereas others showed some loss of activity. In the end, the hydantoin ring and it derivatives were identified as possible alternative scaffolds to an ester group and they could represent interesting bioisosteric replacements for this moiety across other compound series. The loss of activity also comes with the gain in an improved profile of the compound: unlike the ester moiety, these bioisosteres would not be subject in vivo to the action of esterases.

### Table 17.3 SAR resulting from heterocyclic bioisosteres of the ester linkage in tryptophan-derived NK₁ antagonists performed by SparkV10.

<table>
<thead>
<tr>
<th>Index</th>
<th>R₁</th>
<th>R₂</th>
<th>Linkage</th>
<th>Human NK₁ receptor binding affinity ( \alpha ) (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td>1467 ± 450</td>
</tr>
<tr>
<td>2</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td>3.1 ± 1.7</td>
</tr>
<tr>
<td>3</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td>NA</td>
</tr>
</tbody>
</table>

\( \alpha = \frac{X = CH₂}{X = NH} \quad X = O \)
Summary and Conclusions

Bioisosteric replacement is a powerful approach that can suggest modifications for the chemical scaffold of a bioactive compound while retaining the activity and key interaction points. In drug discovery, bioisosteric replacement has a number of potential applications, such as replacement of a patented core scaffold with a different scaffold free from IP restrictions, development of a backup series of compounds, addressing possible metabolic liability or toxicity, and early property exploration of different chemical series for drug development. Scaffold hopping has evolved from just a possible computational application into rapidly expanding complete software tools, which are largely used in the pharmaceutical industry. In this chapter, we have presented a case study of a bioisosteric replacement for NK₁ receptor antagonists: the aim of this investigation was to highlight a successful example of scaffold hopping reported in literature for a ligand-based project and to use it as an example for a retrospective case study. Starting from L-732138 as template molecule as in the published work, the bioisosteric replacement of the ester linkage was performed using commercial software. The aim of this retrospective study was, instead, to highlight the great development, potential, and information available in many of these computational tools for scaffold hopping. Calculations are generally not time demanding, but the many ideas generated from these computational approaches can be of great interest and hopefully of great reward.

References

References


28 Muñoz, M., Rosso, M., and Coveñas, R. (2006) The NK-1 receptor is involved in the antitumoural action of L-733,060 and in the mitogenic action of substance P on human...


62 MacLeod, A.M., Cascieri, M.A., Merchant, K.J., Sadowski, S., Hardwicke, S., Lewis, R.T., MacIntyre, D.E., Metzger, J.M., Fong, T.M., Shepeard, S., Tattersall, F.D.,


Case Study 3: Fragment Hopping to Design Highly Potent and
Selective Neuronal Nitric Oxide Synthase Inhibitors

Haitao Ji and Richard B. Silverman

18.1 Fragment-Based Drug Design

Fragment-based drug design has emerged as an effective tool for lead discovery [1]. The starting point for fragment-based drug design is the identification of low-molecular-weight fragments that bind to a target of interest by fragment screening. The libraries for fragment screening contain hundreds to thousands of small and low-molecular-weight fragments, which are screened at high concentrations [2]. Because the binding affinities of fragments to their target are typically low (in the range of 1 mM to 30 μM), highly sensitive detection methods have been developed to identify the fragments with weak binding affinities [3]. The biophysical techniques for fragment screening include nuclear magnetic resonance (NMR) [4], X-ray crystallography [5], mass spectrometry [6], surface plasmon resonance (SPR) spectroscopy [7], and confocal fluorescence correlation spectroscopy [8,9]. These methods are also sometimes used in a synergistic way [10]. High concentration screening (HCS) is a high-throughput screening (HTS) application to the identification of fragment hits. Due to the low binding affinity to its target, fragments are generally required to be used at high compound concentrations; typically 0.1–2 mM, which is close to the aqueous solubility limit for a substantial number of compounds. In spite of the challenge, HCS is attractive to many fragment-based ligand discovery campaigns because the HTS facilities have been established, HCS is fast and convenient, and its outcome is actually the endpoint of the biological assays (such as $K_i$ or $IC_{50}$). Therefore, HCS can be cost- and resource-effective, provided that the assay platform remains robust and trustworthy. An increasing number of successful examples of HCS have been reported [11].

Another bioassay-based fragment screening technique, called substrate activity screening (SAS), was established to address two key challenges in fragment screening [12]: (i) the accurate and efficient identification of weak binding fragments; (ii) the rapid optimization of the initial weak binding fragments into high-affinity compounds. After fragment hits are identified, the next step is to convert fragment(s) into a lead structure and maintain the drug-like properties of the generated molecule. There are three general strategies for converting
fragments into a drug-like lead structure: fragment evolution, fragment linking, and \textit{in situ} fragment assembly (including dynamic combinatorial chemistry [13], tethering with extenders [14], and \textit{in situ} click chemistry [15]).

There are some internal limitations and problems for those lead generation approaches initiated by fragment screenings. First, fragment screening covers a small fraction of the total chemical diversity space. It is estimated that a collection of $10^3$ fragments can typically sample the chemical diversity space of $10^9$ molecules. Although the combinatorial advantage of fragment screening makes a significant increase relative to HTS, it is still a small fraction of the total diversity space ($10^{60}$) [16]. Second, current fragment screening methods make high demands on the amount, purity, and suitability of target proteins for labeling or crystallization. Small soluble proteins are more amenable for fragment screening. However, for large or membrane proteins (such as G protein-coupled receptors and ion channels), it remains a significant challenge. Third, the geometries and key interactions of the original fragment hits may be altered when they are evolved into lead compounds [17]. Last but maybe the biggest problem, fragment screening only identifies and characterizes fragments for potency, that is, the fragments bind to the “hot spots” (the regions on the protein surface that are the major contributors to ligand free energy of binding). In fact, many binding sites that are responsible for enzyme specificity and/or selectivity are not included in the “hot spots.” The identification of fragments that bind to the sites responsible for enzyme specificity is a crucial step for the discovery of selective ligands. There is a high desire to design lead compounds for a specific enzyme, while leaving related isozymes unaffected.

To meet the above challenges, a wide variety of computational methods have been applied throughout the fragment-based campaigns [18,19]. For example, to address the chemical diversity problem of fragment screening, computational methods have been applied in the design of initial fragment library [20]. \textit{In silico} fragment screening has been used for the large and membrane proteins with which three-dimensional structures are available [21,22]. Structure-based design and molecular docking have been used in the fragment hit-to-lead optimization process. Fragment-based \textit{de novo} design is a computation-based strategy for lead generation in line with fragment screening. Although \textit{de novo} design for lead discovery has been proposed since the 1980s, its success rate has been far behind than those of HTS and virtual screening. The main barriers to \textit{de novo} design are (i) little considerations on the synthetic feasibility of newly constructed structures [23] and (ii) poor predictability of the used scoring functions [24]. On the basis of the concept that the structures of the biologically active compounds for a specific target are discontinuous points in the vast chemical space, scaffold hopping has been proposed to identify the compounds that have similar biological activities but totally different scaffolds [25–27], as described in the other chapters of this book. These methods can decrease the risks of molecular construction or synthetic accessibility, increase the hit rates of lead generation, and offer certain structural diversity. However, in scaffold hopping, the skeleton of the newly designed molecules is confined to the basic architecture of the template structure, which usually comes from a known drug or drug candidate. In addition, mimicking the template
structure using different scaffolds often does not optimize the ligand–receptor
interactions to the maximal extent, because the biologically relevant groups in the
template structures do not offer an optimal match between ligand and receptor.

18.2 Minimal Pharmacophoric Elements and Fragment Hopping

To meet two challenges of current fragment-based approaches in the design of lead
structures: (i) isozyme selectivity and (ii) a better prediction of ligand binding
mode, we recently proposed the concept of minimal pharmacophoric elements
[28]. The minimal pharmacophoric element is smaller than a fragment. It can be an
atom, a cluster of atoms, or a virtual graph or vector(s). The focused fragment
libraries that match the requirement of minimal pharmacophoric elements are
then generated. On the basis of the constructed fragment libraries, various
fragments with different chemotypes, but containing the same minimal pharmaco-
phoric elements, can be derived and a wider chemical space can be explored.
Conversely, in many cases, the region in the active site responsible for ligand
selectivity is rather delicate. Although the other fragment-based approaches can
identify and characterize the fragments located in the “hot spots,” the fragments
that are responsible for isozyme selectivity are generally not located in the “hot
spots.” The mapping of the minimal pharmacophoric elements can preferentially
consider the regions in the active site responsible for isozyme selectivity; thus,
better isozyme selectivity can be incorporated in inhibitor design.

The first step in the determination of the minimal pharmacophoric elements is
the identification of the pharmacophoric sites for a specific drug target by active site
mapping and characterization. If the target structure can be determined by X-ray
crystallography, the multiple solvent crystal structure (MSCS) method can be used
to characterize the potential active site [29,30]. The energetic binding sites on the
protein target for ligand binding can be derived by alanine scanning or site-directed
mutagenesis coupled with biochemical assays. Computational methods including
GRID [31,32], multiple copy simultaneous search (MCSS) [33,34], CS-Map [35], and
FT-Map [36] can also be used to map the active site based on the protein crystal
structures. Monte Carlo simulations, such as grand canonical Monte Carlo
simulation [37], can be used to derive the free energy of binding of the fragments,
rank and cluster fragments. GRID/CPCA is an excellent tool for understanding the
selectivity of inhibitors for a specific target over the other structure-related enzymes
[38]. If the structure of the receptor is unknown, the pharmacophoric sites can be
identified by the structure–activity analysis of ligands and ligand-based pharma-
phore mapping using Catalyst, DISCO, and GASP [39]. Self-organizing maps
(SOM) can be used as a ligand-based tool to predict compound selectivity [40].

Based on the derived minimal pharmacophoric elements, fragment hopping, a
pharmacophore-driven strategy for fragment-based inhibitor design, has been
established to design new bioactive small molecules with novel scaffolds, as shown
in Figure 18.1. On the basis of the derived minimal pharmacophoric elements, two
general-purpose libraries are queried: (i) a basic fragment library, constructed from fragments extracted directly from known drugs and/or drug candidates; (ii) a bioisostere library, constructed based on the bioisosteric principles. The basic fragment library is initially searched to identify all the possible fragments that can match the requirements of the minimal pharmacophoric elements for each pharmacophore. The bioisostere library is then used to generate a focused fragment library with diverse structures. The generated focused fragment library is then interrogated with the rules of metabolic stability and a toxicophore library to yield a focused library for a specific pharmacophore. This focused library is then converted into a LUDI fragment library, and the LUDI program is used to search
the optimal binding position for each fragment of each pharmacophore [41,42]. The next step of this approach is linking these fragments. A side chain library is used for this purpose, in which the synthetic accessibility is considered. SciFinder Scholar [43], in conjunction with the bioisostere library, also plays a key role in securing the synthetic accessibility of the formed chemical bond. The bioisostere library plays an assistant role in enhancing the binding capabilities and optimizing the chemical properties of the generated ligands. The generated ligands are interrogated again with the rules of metabolism stability and the toxicophore library. The ligand molecules generated by this iterative process are then docked into the active site, scored with the consensus scoring functions [44], and filtered with absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) considerations [45,46]. If the ligands generated are not satisfactory, the molecule is reconstructed using the generated focused fragment libraries, the side chain library, and the bioisostere library. If the generated ligand is interesting, it will be synthesized and biologically tested.

The core of fragment hopping is the derivation of the minimal pharmacophoric elements for each pharmacophore. Fragments are then generated to match the requirement of minimal pharmacophoric elements based on the basic fragment and bioisostere libraries. Fragment hopping is an open system that effectively integrates other techniques and provides a more efficient pathway to generate not only potent but also selective inhibitors. This approach has been successfully applied to the design of potent and selective inhibitors for neuronal nitric oxide synthase (nNOS).

18.3 Fragment Hopping to Design Novel Inhibitors for Neuronal Nitric Oxide Synthase

Nitric oxide (NO) is synthesized by a family of enzymes called nitric oxide synthase (NOS, EC 1.14.13.39), which catalyzes an NADPH- and O2-dependent five-electron oxidation of L-arginine to L-citrulline and NO via intermediate N\(^{G} \)-hydroxy-L-arginine [47]. Three distinct isozymes of NOS have been identified in mammals [48], and each is associated with a different physiological function: nNOS, which generates NO in the CNS, is involved in neurotransmission and long-term potentiation; endothelial nitric oxide synthase (eNOS)-derived NO is involved in the regulation of smooth muscle relaxation and vascular tone; and a third, inducible isoform in macrophages (iNOS), is important in the immune system defense against pathogens and tumor cells [49]. The sequence homology of these three mammalian NOSs is 50–60%.

NOS is active as a homodimer with each subunit containing a C-terminal reductase domain (with the binding sites for NADPH, FAD, and FMN) and a N-terminal oxygenase domain containing the heme prosthetic group. The substrate L-arginine and a redox cofactor, \((6R)-5,6,7,8\)-tetrahydro-1-biopterin (H\(_4\)B), both bind near the heme center in the oxygenase domain [50]. nNOS and eNOS are constitutively expressed and intermittently produce small amounts of NO. In contrast, iNOS is inducible by cytokines and produces a large amount of NO for
both a cytoprotective and a cytotoxic effect. The constitutive enzymes bind calmodulin (CaM) reversibly in a Ca^{2+}-dependent manner, and their activity is regulated by local Ca^{2+} concentration, while iNOS does not depend on the intracellular calcium level because iNOS carries CaM as a permanently bound subunit. Electrons are transferred from NADPH via FAD and FMN to the heme of the other subunit. This flow of electrons during catalysis occurs from the reductase domain of one monomer to the oxygenase domain of the other monomer, so an intact homodimeric form of the enzyme is essential for full enzyme activity.

Overproduction of NO from nNOS has been associated with harmful effects in the central nervous system, including stroke [51], Alzheimer’s disease, Parkinson’s disease, schizophrenia, AIDS dementia, and cerebral palsy [52]. Radical nitric oxide (NO\(^{•}\)) can react rapidly with superoxide (O\(_{2}^{•−}\)) in aqueous media to generate the highly reactive peroxynitrite anion (ONOO\(^{−}\)). Either nitric oxide or peroxynitrite causes oxidative damage, protein nitration, and, therefore, leads to neurodegeneration [53]. NO overproduction by iNOS is critical for the immune response [54]. NO produced by eNOS in endothelial cells has mainly physiological roles, such as maintaining normal blood pressure and flow [55–57]. Accordingly, inhibition of nNOS, but not of iNOS and eNOS, could provide an effective therapeutic approach [58–60]. On the other hand, selective inhibitors could also be useful tools for investigating other biological functions of NO [61].

The crystal structures of the dimeric oxygenase domain for all three NOS isozymes have been solved [62–65], which provide a possibility for structure-based inhibitor design. However, the design of selective inhibitors has proven to be a challenging problem because the active sites of NOS isozymes are highly conserved. Out of 18 residues within 6 Å of the substrate binding site, 16 are identical, and the side chain of one of these two dissimilar amino acids points out of the substrate binding site [66]. The other two key challenges in the designing of nNOS inhibitors are (i) the very large active site of nNOS, 1000–1100 Å\(^3\) that makes it difficult to design an inhibitor that can contact the protein surface well enough to produce good binding, and (ii) cell permeability: 18 of 30 residues/cofactor side chains that point into the active site of NOS are polar or charged. Clusters of acidic residues/cofactor side chains, especially residues E592 and D597, the heme propionate groups, and two low pK\(_{a}\) polar side chains (residues Y562 and Y588), form a strongly acidic local environment. This environment requires the inhibitor to contain positively charged electrostatic or hydrogen bond donor groups, such as amines, which creates a problem for the design of a cell-permeable inhibitor [67].

Previously, the nitroarginine-containing dipeptide or peptidomimetic inhibitors were identified as a family of compounds that had high potency and dual selectivity for inhibition of nNOS over eNOS and iNOS. The first three potent nNOS inhibitors discovered from these series were 1 [68], 2 [69], and 3 [70] (Scheme 18.1).

The selectivity of the dipeptide/peptidomimetic inhibitors for nNOS over eNOS was investigated by crystallographic analysis [71–73], which indicates that a single-residue difference in the active site, rat nNOS D597 versus bovine eNOS N368, accounts for a majority of the selectivity for nNOS over eNOS [71]; this high selectivity is determined by the α-amino groups of the inhibitors. The maximum
electrostatic stabilization arising from residues D597 and E592 and the α-amino group of inhibitors 1–3 forces inhibitors to adopt a curled conformation. Such stabilization is rather weak in eNOS because N368 does not bear a negative charge. Inhibitors 1–3 in eNOS adopt an extended conformation, and the α-amino group is shifted away from the corresponding selective region defined in nNOS [71]. The minimal pharmacophoric elements are extracted on the basis of the above active site analyses [28,74] and inhibitor structure–activity relationship studies [58,59], as shown in Figure 18.2. An amidino group is positioned in the same place as the guanidino group of substrate L-arginine. This group is the minimal binding element to form a charge–charge interaction and H-bonds with the carboxylate side chain of E592 and the backbone amide of W587. One sp3-hybridized nitrogen cation is placed in a selective region defined by D597 of nNOS and N368 of eNOS. The other three nitrogen atoms are placed close to the heme propionate to form a charge–charge interaction and H-bonds. In the S pocket, steric and hydrophobic effects play important roles in ligand binding. The steric effect is prominent at the position close to D597 and the heme propionate, as indicated by the circles in Figure 18.2.

A focused fragment library was generated based on the proposed minimal pharmacophoric elements for each pharmacophore. The fragments were then docked into the active site of nNOS where the corresponding pharmacophore is located. It is noteworthy that the amidino group and the nitrogen atoms are directional and require rather rigorous positioning for optimal ligand–receptor interactions. Thus, the fragments that are able to match their requirements are limited. However, the fragment options for hydrophobic and steric interactions are rather broad when the basic fragment library and the bioisostere library are

<table>
<thead>
<tr>
<th>Compound</th>
<th>( K_i ) (µM)</th>
<th>Selectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nNOS</td>
<td>cNOS</td>
</tr>
<tr>
<td>1</td>
<td>0.13</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>314</td>
</tr>
<tr>
<td>3</td>
<td>0.10</td>
<td>128</td>
</tr>
</tbody>
</table>

Scheme 18.1 Chemical structures of 1–3 and their inhibitory activities toward three isozymes of nitric oxide synthase (NOS).
interrogated. That is, targeting hydrophobic and steric interactions would offer more diverse fragments for each pharmacophore initially. A small subset of this focused fragment library is described below. To match the requirements of the amidino group and hydrophobic/steric effects, the 2-aminopyridine group was selected as a basic fragment. One advantage of the 2-aminopyridine fragment is that the $pK_a$ value of 2-amino-6-methylpyridine is 6.69 [75]. This fragment could act as a charge switch: in the small intestine, the fragment could be in its neutral form, which is favorable for absorption; in the NOS active site, the local acidic environment could convert it into the positively charged form, which is favorable for binding. Starting with the nitrogen atom close to D597 in Figure 18.1, the pyrrolidino fragment was generated as a substitute for the $\alpha$-amino group of 1–3. The pyrrolidino group can not only meet the charge–charge interaction requirement for nNOS selectivity but also match the steric effect requirement for NOS binding. Another advantage of using the pyrrolidine ring is that the secondary amino group is more lipophilic and has less polar surface area (PSA) compared to the primary amino group of 1–3, which is better for in vivo inhibitor delivery [76]. The ethylenediamine fragment was chosen to form a charge–charge interaction and H-bonds with the two heme propionate groups. After the linking of these fragments, compound 4 in Scheme 18.2 emerged as the desired molecule.

Figure 18.2 Minimal pharmacophoric elements for selective nNOS inhibitors. An amidino group is positioned close to E592. A yellow nitrogen atom is close to D597. The regions where hydrophobic and/or steric interactions play important roles are indicated by circles. Three blue nitrogen atoms are placed close to the heme propionate. (Reprinted with permission from Ref. [28]. Copyright 2008, American Chemical Society.)
NOS enzyme inhibition assays show that the nNOS inhibitory activity ($K_i$) for 4 is 388 nM with high selectivity for nNOS over eNOS (1100-fold). Compound 4 is a racemic mixture; theoretically, only the active ($3'S,4'S$)-enantiomer binds to the active site, which means that the single enantiomer should generate enhanced inhibitory activity. The selectivity of 4 between nNOS and iNOS is 150-fold. These selectivities are comparable to those attained with nitroarginine-containing dipeptide derivatives 1–3 [28]. The crystallographic analysis shows that only the designed enantiomer was bound to the active site (PDB id, 3B3N) [77]. Figure 18.3a shows a superimposition of the binding conformations and the predicted bioactive

Figure 18.3 (a) Superimposition of the binding conformation (blue) and predicted bioactive conformation (yellow) of 4 in the active site of rat nNOS. (b) Superimposition of the binding conformations of 4 (blue, PDB id, 3B3N) and 2 (yellow, PDB id, 1P6I) in rat nNOS. The heme (orange), $H_4B$ (violet), and structural water (green) involved in the binding of 4 to nNOS are shown. The distances of some important H-bonds between the residues, structural water, cofactors, and inhibitors are given in Ångstroms (Å). (Reprinted with permission from Ref. [28]. Copyright 2008, American Chemical Society.)
conformation of 4 with nNOS, and Figure 18.3b shows a superimposition of the binding conformation of 2 (PDB id, 1P6I [71]) and 4 with nNOS. Both figures indicate that fragment hopping can be used to generate new lead structure that mimics the mode of action of the peptide or peptidomimetic inhibitors, thus providing a new peptidomimetic strategy.

18.4 Fragment Hopping to Optimize Neuronal Nitric Oxide Synthase Inhibitors

The fragment hopping strategy can not only generate new lead structure but also facilitate the structural optimization process [78]. On the basis of the structure of 4, fragment hopping effectively extracted new pharmacophoric elements to improve inhibitor drug-like properties and to generate appropriate fragments for lead optimization. One small cavity lined with the side chains of P565, A566, V567, and F584 in the S pocket. A methyl group was introduced to the para-position of the 2-aminopyridine ring. Fragment hopping also identified a larger hydrophobic pocket lined with M336, L337, Y706, and W306 (from the other subunit). Halogen-substituted phenylalkyl fragments were introduced at the terminal amino group of the ethylenediamine side chain of 4. It is worth noting that this fragment explores a new binding site compared to inhibitors 1–3. The halogen-substituted phenyl groups were used for three reasons: (i) The hydrophobic phenyl group can be stabilized in a shallow hydrophobic pocket defined by M336, L337, Y706, and W306 (from the other subunit). The residue that corresponds to rat nNOS L337 is N115 in murine iNOS or T121 in human iNOS. The hydrophobic halogen-substituted phenylalkyl group could be used to differentiate the hydrophobic L337 in nNOS from the polar N115 and T121 residues in murine and human iNOS, respectively. Therefore, this scaffold enables the design of selective nNOS inhibitors over not only eNOS but also iNOS. (ii) The introduction of the phenylalkyl group renders the amino group of the ethylenediamine fragment a secondary amine, which decreases its polar surface area, making it much more favorable for biomembrane permeability compared to a primary amino group. (iii) The introduction of halogen atoms at the para- or meta-positions of the phenyl group blocks/decreases metabolic degradation of the phenyl group as well as increases its lipophilicity.

After the syntheses and bioassays of 26 derivatives, compounds 5 and 6 were identified. Compared to 4, compounds 5 and 6 exhibit greater nNOS inhibitory activity and possess similar or better nNOS/eNOS selectivity. The crystal structure of nNOS in complex with 5 showed a bioactive conformation of 5 as predicted and mimic the mode of action of the dipeptide nNOS inhibitors. As expected from the chirality of the active site, only one of the cis-enantiomers of 5, the (3'S,4'S)-isomer, was bound to the active site, as shown in Figure 18.4a (PDB id, 3B3O). The binding conformation of 5 in the crystal structure is very similar to that predicted in fragment hopping (Figure 18.4b) and mimics the binding modes of 1–3 (Figure 18.4c). The 2-amino-4-methylpyridine group of 5 forms two H-bonds and electrostatic interactions with the carboxylic acid group of E592, just as the
nitroguanidino groups of 1–3 do. The nitrogen atom of the pyrrolidine ring of 5 forms a direct electrostatic interaction with E592 and is involved in a H-bond network with the carboxylic acid group of D597 via two structural water molecules (Figure 18.4b), just as the α-amino groups of 1–3 do (Figure 18.4c). The 4-methyl group is located in a small hydrophobic/steric pocket defined by the side chains of V567 and F584, as it was designed for. The NH group of the ethylenediamine fragment of 5 that is attached to the pyrrolidine ring forms a hydrogen bond to the heme propionate group, just as the amido NH groups of 1 and 3 and the secondary amino group of 2 do. The other nitrogen atom of the ethylenediamine fragment of 5 is involved in electrostatic interactions with heme propionate groups. The 4-chlorobenzyl group of 5 is close to hydrophobic residues M336, L337, and W306 (from the other subunit). Compared to the dipeptide or peptidomimetic inhibitors, compounds 5 and 6 have a better predicted biomembrane permeability, metabolic stability, and bioavailability, which suggests that 5 and 6 were evolved well from 1–3 and can have a better activities in vivo.

18.5 Application of Neuronal Nitric Oxide Synthase Inhibitors to the Prevention of Cerebral Palsy

Cerebral palsy is one of the most severe consequences of hypoxia-ischemia (HI) before birth and is common in premature infants, with 750 000 persons affected in the United States [79]. It has one of the highest indices of disease burden with direct effects on individual, family, and social institutions (annual cost $8.2 billion) [80] that last the entire lifetime. There is no known treatment to protect the fetus from hypoxic brain injury leading to cerebral palsy [81], despite a reduction in the mortality of high-risk infants [82]. The overproduction of NO by nNOS has been strongly implicated in the subsequent development of cerebral palsy in premature [83] and full-term infants [84,85]. The administration of nNOS-selective inhibitors may provide a new preventive strategy. Maternal administration of 5 and 6 leads to
inhibition of fetal brain NOS activity and a decrease in the production of fetal brain NO (Figure 18.5) [66]. Because of the high selectivity of 5 and 6 for nNOS relative to eNOS (1000- and 2000-fold, respectively), there are no detrimental cardiovascular effects. Neither the heart rate nor the blood pressures of the dams were affected by these compounds, unlike the cardiovascular effects observed with poorly selective compounds, such as L-N\(^{\text{O}}\)-nitroarginine methyl ester (L-NAME) and 7-nitroindazole.

A truly striking difference was observed in the neurobehavioral assessment of kits born to 5- and 6-treated dams relative to those from the saline control dams. The P1 kits from saline-treated dams had a large incidence of discovered fetal/neonatal deaths (47% or 16/34) and severe neurobehavioral abnormalities (35% or 12/34) compared to those from dams treated with 5 or 6 [none (0/49) died and only 14% (7/49) had severe neurobehavioral abnormalities] (Figure 18.6). Only 9% (3/34) of the kits from saline-treated dams were born normal, whereas 76% (37/49)
of the kits born to 5- and 6-treated dams were born normal (in two litters, all 19 of the kits were normal). None of the compounds caused any detectable systemic toxicity in the rabbit dams. The availability of these potent inhibitors enables a preventive strategy to be employed in high-risk mothers to ameliorate the pathophysiological cascades resulting in postnatal cerebral palsy.

Acknowledgment

The authors are grateful for financial support from the National Institutes of Health Grant GM 49725. HJ thanks the Pulmonary Fibrosis Foundation I. M. Rosenzweig Young Investigator Award (Award number, 235170) for the financial support.

References

Case Study 3: Fragment Hopping to Design Highly Potent and Selective


294 18 Case Study 3: Fragment Hopping to Design Highly Potent and Selective


References


77 Igarashi, J., Li, H., Jamal, J., Ji, H., Fang, J., Lawton, G.R., Silverman, R.B., and


Index

a
absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) 39, 283
Accelrys FCFP 4 fingerprint 110
activity Cliff-forming scaffolds 71, 72
– activity Cliff concept 71
– multitarget Cliff-forming scaffolds 71–73
AllChem 94
N-alpha-(2-naphthyl-sulphonyl-glycyl)-Dl-pamidinophenylalanyl-piperidine (NAPAP) ligand 238, 239
amino acid
– interaction types 223
– pharmacophoric properties 221
– residues, responsible for binding of 261
– sequence 172, 260
– specific binding 223
AstraZeneca 17, 87
– virtual library 95
acyclic fragments 233
AZ-VL system 97
calcium-dependent NOS activity 290
calmodulin (CaM) 284
Cambridge Structural Database (CSD) 271
CAS. see Chemical Abstracts Service (CAS)
CAS (Chemical Abstracts Service) registry 43, 46
– nonredundant library 46
CATS. see chemically advanced template search (CATS)
CATS 2D autocorrelation function 111
CATS fingerprints 108, 111, 114
cavity-independent fingerprints (APIF) 222
cerebral palsy, prevention 289
CG. see chemical graph (CG)
ChEMBL database 44, 49, 52, 53, 55, 62, 97
ChEMBL data set 111
Chemical Abstracts Service (CAS) 16, 43
chemical cliché 44
chemical graph (CG) 131, 133
– paracetamol molecular structure 132
chemically advanced template search (CATS) 119–122
– CATS2 atom types 121
– descriptor
– generation 122
– median BEDROC performance 125
– fuzzy molecular representations 120
– modify ATS descriptor 120
chemical patent 15, 17, 28, 34, 36
chemical space 83, 84
– comprehensive enumeration 85
– BOOMSLANG algorithm 88
– fragments 85, 86
– Kohonen map 86
– reagents 87, 88
– ring systems 86, 87
– SMIRKS manipulation 87

b
bacterial neuraminidase, X-ray crystal structure 241
BCI fingerprint 110
benzimidazole scaffold 3
BI CLAIM system 97
binding affinity 3, 48, 221, 222, 264, 265, 268, 269, 273
binding conformation, superimposition of 287
BindingDB 64
bioisostere encoding 144
bioisosteric replacements 247, 259
BIOSTER database 144
bis(trifluoromethyl)phenyl rings 266
BROOD system 179
– bioisosteric similarity methods 180

© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.
VEHICLE 86
iterative stochastic search 85
chemoinformatics 107
claimed intellectual property 17
COBRA database 111, 114, 123
community-selective scaffolds 69
computed molecular properties 107
carbohydrate recognition 4
calcium-calmodulin-dependent protein kinase II (CaMKII) 277
case study 99
cassette library 252
cationic amino groups 17
C atoms 236
cyclooxygenase-2 (COX-2) 235
dati data visualization 55
Daylight fingerprints (DFPs) 108, 110, 138, 139, 142
diarylthiazolotriazole 4
database-based approach 108
2,5-dioxopiperazine 267
DOGS program 94
DrugBank 43
drug discovery
application of Spark 206, 207
automatic devices 99
and design 3
Drug Guru 36
similarity searching in 155–157
Drug Guru program 36, 89, 90, 99
drug metabolism and pharmacokinetic (DMPK) 247
3D scaffold replacement method 175
BROOD 179, 180
ChEMBLstructures 177
commercial structure 177
CSD structures 176
databases 176, 177
filtering 177, 178
fragment generation 177, 178
fragment replacement search 178
molecule databases 175
ReCore 179
scoring 178
SHOP, as novel replacement scaffolds 178, 179
virtual structures 177
workflow 175, 176
2D structure 98
ECFP fingerprints 111, 114
electronic laboratory notebooks (ELNs) 95
enantiospecific inhibition 252
endothelial nitric oxide synthase (eNOS)-derived NO 283
ER activity, designed compounds with 236
ester linkage, in tryptophan-derived NK1 antagonists 269
estrogen receptor (ER) 241
affinity 235
site point constraints 240
ethylendiamine fragment 286, 288
Extended Connectivity Record (ECTR), programs 16
eXtended electron distribution (XED) force field. see XED force field
feature point pharmacophores (FEPOPS) 157–159
active recall, and scaffold hopping descriptor assessment 166
alignment of protein kinase C inhibitor 165
alternative alignment 168–170
atomic properties tocentroids, assignment of 161
average similarity of probes 168
chemical space uniqueness 171, 172
computation 159–162
defining scaffold uniqueness 167
descriptor 160
generation of feature point pharmacophores 161
k-means and k-medoids calculations 160, 162
clustering 158
multiscale reduction of ATP atoms 159
multiple flexible conformers 160
Omega software, use of 162
perspective 172, 173
recovery of validated HTS actives, clusters, scaffolds, and ring systems 167
scaling and correlations 162
in silico target prediction 170, 171
in similarity searching and scaffold hopping 164–168
<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>– correspondence between MIL file and</td>
<td>21, 22</td>
</tr>
<tr>
<td>– economic importance</td>
<td>16</td>
</tr>
<tr>
<td>– encoding</td>
<td>18</td>
</tr>
<tr>
<td>– exact R groups</td>
<td>19</td>
</tr>
<tr>
<td>– fused R groups</td>
<td>19, 20</td>
</tr>
<tr>
<td>– inexact R groups</td>
<td>19</td>
</tr>
<tr>
<td>– r_group records</td>
<td>19</td>
</tr>
<tr>
<td>– Menguin program</td>
<td>20, 21</td>
</tr>
<tr>
<td>MARPAT systems</td>
<td>17</td>
</tr>
<tr>
<td>maximum common substructure (MCS)</td>
<td>9–11, 44, 84, 91</td>
</tr>
<tr>
<td>– algorithms, challenges</td>
<td>9</td>
</tr>
<tr>
<td>MDDR database</td>
<td>109</td>
</tr>
<tr>
<td>– activity classes from</td>
<td>109</td>
</tr>
<tr>
<td>– scaffold-hopping searches</td>
<td>110</td>
</tr>
<tr>
<td>MDDR data set</td>
<td>139, 142</td>
</tr>
<tr>
<td>MDL fingerprint</td>
<td>108, 110, 111</td>
</tr>
<tr>
<td>medicinal chemistry scaffolds, enumerated</td>
<td>47</td>
</tr>
<tr>
<td>– possibilities</td>
<td>47</td>
</tr>
<tr>
<td>– integer descriptors</td>
<td>46</td>
</tr>
<tr>
<td>– Lipkus’ topology analysis</td>
<td>47</td>
</tr>
<tr>
<td>– nonredundant library, in CAS registry</td>
<td>46</td>
</tr>
<tr>
<td>– ring systems</td>
<td>46, 47</td>
</tr>
<tr>
<td>– topologies of voids</td>
<td>46, 47</td>
</tr>
<tr>
<td>– VEHICLE database</td>
<td>47</td>
</tr>
<tr>
<td>MEGx database</td>
<td>114</td>
</tr>
<tr>
<td>Menguin program</td>
<td>18</td>
</tr>
<tr>
<td>Merged Markush Service (MMS)</td>
<td>16, 17</td>
</tr>
<tr>
<td>MFP2 fingerprint</td>
<td>112</td>
</tr>
<tr>
<td>MIL file</td>
<td>18</td>
</tr>
<tr>
<td>molecular cloning studies</td>
<td>248</td>
</tr>
<tr>
<td>molecular geometry</td>
<td>107</td>
</tr>
<tr>
<td>molecular interaction fingerprints</td>
<td></td>
</tr>
<tr>
<td>– ligand-annotated target</td>
<td>217–219</td>
</tr>
<tr>
<td>– protein–ligand complexes</td>
<td></td>
</tr>
<tr>
<td>– three-dimensional (3D) structures</td>
<td>215</td>
</tr>
<tr>
<td>– protein-ligand pharmacophores</td>
<td>217</td>
</tr>
<tr>
<td>– target-annotated ligand</td>
<td>215, 216</td>
</tr>
<tr>
<td>– interacting atom/fragment</td>
<td>216, 217</td>
</tr>
<tr>
<td>– true target-ligand</td>
<td>220</td>
</tr>
<tr>
<td>– association of</td>
<td>220–222</td>
</tr>
<tr>
<td>– interaction pattern</td>
<td>222–225</td>
</tr>
<tr>
<td>molecular libraries production center</td>
<td>248</td>
</tr>
<tr>
<td>network (MLPCN)</td>
<td></td>
</tr>
<tr>
<td>molecular libraries screening center</td>
<td>248</td>
</tr>
<tr>
<td>network (MLSCN)</td>
<td></td>
</tr>
<tr>
<td>molecular topology</td>
<td>107</td>
</tr>
<tr>
<td>molecular weight</td>
<td>84–86, 88, 107, 261, 279</td>
</tr>
<tr>
<td>molecule-based approach</td>
<td>108</td>
</tr>
<tr>
<td>Molecule Evoluator</td>
<td>89</td>
</tr>
<tr>
<td>MotifScore</td>
<td>223</td>
</tr>
<tr>
<td>multiple objective optimization</td>
<td>90</td>
</tr>
<tr>
<td>multiple solvent crystal structure (MSCS)</td>
<td>281</td>
</tr>
<tr>
<td>method</td>
<td></td>
</tr>
<tr>
<td>Murcko scaffold</td>
<td>110</td>
</tr>
<tr>
<td>MUV data set</td>
<td>111</td>
</tr>
<tr>
<td>n natural products, source of novel scaffolds</td>
<td>45, 46</td>
</tr>
<tr>
<td>NATx database</td>
<td>114</td>
</tr>
<tr>
<td>neurokinin 1 (NK1) receptor</td>
<td>260</td>
</tr>
<tr>
<td>– antagonists</td>
<td>262</td>
</tr>
<tr>
<td>– residues, involved in binding of nonpeptide antagonists</td>
<td>265</td>
</tr>
<tr>
<td>neuronal nitric oxide synthase (nNOS)</td>
<td>283</td>
</tr>
<tr>
<td>neuropeptide SP, NMR structure</td>
<td>265</td>
</tr>
<tr>
<td>neuropeptide substance</td>
<td>261</td>
</tr>
<tr>
<td>nitric oxide synthase (NOS), chemical</td>
<td>285</td>
</tr>
<tr>
<td>structures</td>
<td></td>
</tr>
<tr>
<td>nitroarginine-containing dipeptide/</td>
<td></td>
</tr>
<tr>
<td>peptidomimetic inhibitors</td>
<td>284</td>
</tr>
<tr>
<td>nNOS</td>
<td></td>
</tr>
<tr>
<td>– binding conformation</td>
<td>289</td>
</tr>
<tr>
<td>– crystal structure</td>
<td>288</td>
</tr>
<tr>
<td>– dipeptide/peptidomimetic inhibitors</td>
<td>284</td>
</tr>
<tr>
<td>– inhibitors, minimal pharmacophoric</td>
<td>286</td>
</tr>
<tr>
<td>elements</td>
<td></td>
</tr>
<tr>
<td>N5SB polymerase</td>
<td>3</td>
</tr>
<tr>
<td>NVP-AUY922, HSP90 inhibitor</td>
<td>6</td>
</tr>
<tr>
<td>o objective and invariant scaffold</td>
<td>7</td>
</tr>
<tr>
<td>representations</td>
<td></td>
</tr>
<tr>
<td>– molecular frameworks</td>
<td>7, 8</td>
</tr>
<tr>
<td>– scaffold tree</td>
<td>8</td>
</tr>
<tr>
<td>oxazolidinedione</td>
<td>268</td>
</tr>
<tr>
<td>2-oxopiperazine</td>
<td>267</td>
</tr>
<tr>
<td>p ParaFit program</td>
<td>188–190</td>
</tr>
<tr>
<td>ParaSurf program</td>
<td>188–190</td>
</tr>
<tr>
<td>PGVL system</td>
<td>97</td>
</tr>
<tr>
<td>pharmacological similarity</td>
<td>195</td>
</tr>
<tr>
<td>– chemotype trap</td>
<td>195, 196</td>
</tr>
<tr>
<td>– G protein-coupled receptors</td>
<td>195</td>
</tr>
<tr>
<td>– ligand</td>
<td>195, 198</td>
</tr>
<tr>
<td>– pattern 196–199</td>
<td></td>
</tr>
<tr>
<td>– shape compatibility</td>
<td>195</td>
</tr>
<tr>
<td>– van der Waals forces</td>
<td>195</td>
</tr>
<tr>
<td>pharmacophore-based interacting</td>
<td></td>
</tr>
<tr>
<td>fingerprints</td>
<td>217</td>
</tr>
<tr>
<td>pi–cation interaction</td>
<td>219</td>
</tr>
<tr>
<td>Pipeline Pilot software</td>
<td>110</td>
</tr>
<tr>
<td>piperazine ring system</td>
<td>268</td>
</tr>
</tbody>
</table>
polarity 3
polar surface area (PSA) 107, 286
potency-based scaffold discontinuity score (PScS) 71
privileged scaffolds 11
privileged substructures 63
Pro271 265
programmable compounds 99
promiscuous chemotypes 68
promiscuous scaffolds, in drugs 70, 71
protein and ligand fingerprints (PLFPs) 221
protein–ligand association fingerprint 220
protein ligand interaction explorer (PROLIX) 223
protein–ligand interaction fingerprints 218
protein–ligand interaction pharmacophores 217
protein–ligand structural alignments 222
pseudo-SMILES encoding 138
PubChem 32, 33, 43, 62, 86
Putative ligand structures 232
PyMOL graphical user interface 232
pyrrolidine ring 286

q
quantitative structure–activity relationship (QSAR) models 152
– feature tree representations of molecules 152
– Free–Wilson analysis 33
– inverse 89, 90, 98

r
rabbit kits
– neurobehavioral assessment 290, 291
– nitric oxide synthase/nitric oxide concentration 290
random virtual compounds 83
RDKit fingerprint 124
reaction SMARTS 93, 94
readily synthesizable compounds 94
– construction 94, 95
– iterative approaches 96
– outside big Pharma 96
– searching 95
real-life reactions, for in silico molecule construction 93, 94
ReCore program 179
reduced graph (RG) 131, 133, 141
– comparison and usage 137, 138
– augmenting fingerprints, with edit distance 140, 141
– bioisostere encoding 144
– conventional fingerprinting 138, 139
– dopaminergic activity retrieved 138
– extended reduced graph fingerprints 141–143
– graph matching approaches 143, 144
– RG-specific fingerprints 139, 140
– structural diversity of 5HT reuptake inhibitors 142
– key challenges, to assess molecular similarity 145
reference structure 107
Reflex method 240
R/F reduction scheme 135
RG. see reduced graph (RG)
ring fragments 233
rotating spherical polar fourier expansions 185, 186

s
SAR data 97
scaffold composition
– drugs and small-molecule libraries 45, 46
– to interpret bioactivity data 48
– SCINS code 48
– structure–activity relationships 48
– medicinal chemistry libraries 46, 47
– medicinal chemistry space 41–45
– molecular framework analysis 45, 46
– drug and drug-like library 46
– natural product library 46
– study based on MCS 44
scaffold definition 64, 65, 84
scaffold diversity
– defining 41
– in medicinal chemistry 39–41
– medicinal chemistry space
– metrics for quantifying 48–53
– visualizing scaffold diversity 53–56
– of pharmaceutical targets 76–78
– structural relationships between scaffolds 76, 77
scaffold hopping 3, 31, 76, 78, 107, 240–242, 247–248
– access novel calcium T-type channel inhibitors 250–253
– access novel glycine transporter type 1 (GlyT1) inhibitors 253–254
– application of 235
– bioisostere 259
– capability 110
– from a CCR3 antagonist to existing library compound 173
– into chiral 3-aminopiperidine cores 251
– into [3.1.0] cores 252
– de novo design 231
– Drug Guru program 36
– fast-follower project 36
– fragment-based drug design 279–281
– minimal pharmacophoric elements/
  fragment hopping 281–283
– neuronal nitric oxide synthase 283–288
– optimize neuronal nitric oxide synthase
  inhibitors 288–289
– Free–Wilson analysis 33–35
– neurokinin 1 receptor (NK1R) 260, 261
  – target active site and binding mode 264,
    265
– neurokinin 1 receptor antagonists 261–264
  – bioisosteric replacements 266–273
– neurokinin 1 (NK1) therapeutic areas 259,
  260
– neuronal nitric oxide synthase inhibitors,
  application of
  – cerebral palsy prevention 289–291
  – periscope system assisting 32
  – potential 76
– CATS1 and CATS2 descriptors 125
– descriptors, actives-retrieving performance
  categories 124, 125
  – determining 125
  – heatmap depicting p-values 124
  – retrospective evaluation of enrichment
    and 122–126
  – ROC-related BEDROC score 123
  – virtual screening benchmark 123
  – predictive studies, using 2D
    fingerprints 112–114
  – prospective applications 126–128
  – substructure searching 32, 33
– T-type calcium channel inhibitors 248–250
– using 2D fingerprints, retrospective
  studies 109–112
– in virtual screening 78
scaffold network graphs, of drugs 56
scaffold representations
– history 4–7
– of kinase inhibitor lapatinib 40
– libraries A and B 42
scaffold space, visualizations 96, 97
scaffolds, with defined activity progression 73
– activity profile sequences 73–75
– conserved scaffolds 75
scaffold-target family profiles 70
scaling bit positions 217
SCINS (scaffold identification and naming
  systems) 48, 52, 163, 164
– descriptor 163, 172
ScreenScore 234
search algorithm 22–25
– i3am search algorithm 23
– matching R groups 25
  – exact R groups 25–27
  – fused R groups 28
  – hydrogen atoms 29, 30
  – inexact R Groups 27, 28
  – managing multiple fragment/R group
    matches 30, 31
– molecules matched against Markush 24, 25
  – indole core 24
  – matche C2–6 alkoxy 25
  – methoxy carbonyl and chloride groups 25
  – R groups and portions of molecule
    unmatched 24
  – total score 25
  – Xenon atoms 24
selectivity, of scaffolds 63
– privileged substructures 63, 64
– target community-selective scaffolds 64–66
– target-selective scaffolds 67
self-organizing maps (SOMs) 54, 126, 281
Shannon entropy 51
shell residue fingerprint (SRFP) 223
SHOP program 178, 179
similarity searching, in drug discovery
  155–157
  – 4D chemical similarity method (see feature
    point pharmacophores (FEPOPS))
  – 2D similarity methods 156, 157
  – graph-like representations, of
    compounds 156
  – high throughput screening (HTS) 155
  – ligand-based hypothesis 155
– Markush structures 156
– scaffold hopping 156
  – structure–activity relationships 156
Skelgen 231
– cyclin dependent kinase 2 (CDK2) 235
– cyclooxygenase-2 (COX-2) 235
– estrogen receptor 235, 236
– histamine H3 inverse agonists 236
– principles 231
– PyMOL graphical user interface 232
– retrospective validation study 235
– scaffold hopping 240–242
  – using fixed fragments 237, 238
  – using site points 238–240
– structure generation/optimization
  – fragments/fragment sets 232–233
  – ligand-based design 234–235
Index | 303

--- scoring 234
– validation studies 235
Skelgen fragment 237
SMIRKS 87, 89
SOMs. see self-organizing maps (SOMs)
Spark
– in drug discovery scenarios, applications 206, 207
– P38 kinase inhibitor fragment growing using 207–209
– workung 202
– selecting fragment replacement site 206
– software 202, 205
SparkV10 270, 271, 272
– heterocyclic bioisosteres of the ester linkage 273
– template molecule L-732138 271
spherical harmonic surface
– shape similarity 186, 187
spherical harmonic surfaces 183–185
stearoyl-CoA desaturase inhibitor 3
steric interactions 272, 285
structural interaction fingerprint (SIFt) 218
structure–activity relationships (SARs) 61, 84
structure-based de novo design 90–92
substance P receptor (SPR). see neurokinin 1 receptor
substrate activity screening (SAS) 279
Surflex-Dock score 225
SVM model 221
SYNOPSIS tool 93

\[ t \]
tachykinin NK\(_1\) family 264
tachykinin receptor 1 (TACR1). see neurokinin 1 receptor
Tanimoto coefficient 107, 108
– values 78
tan ribbons 216
target-annotated ligand fingerprints 216
target community-selective scaffolds 64–66
target–ligand fingerprints 220
target promiscuity of scaffolds 67
– promiscuous BM scaffolds and CSKs 67–70
– promiscuous scaffolds in drugs 70, 71
– scaffold–target family profiles 70
target selectivity (TS) value 67
thiophene S-oxide 4
TIFP fingerprint 224
topological structure, autocorrelation 119

toxicity 3
tree maps 54
Tripos Unity fingerprint 110
tryptophan-derived NK\(_1\) receptor antagonists 266, 270
– binding affinity 273
– ester linkages, case study 266, 273
– heterocyclic bioisosteres 267
T-type Ca\(^{2+}\) inhibitors
– high-throughput screen (HTS) 249
– Merck piperidine series 251
– second-generation 249
T-type calcium channel family 248
Tversky coefficient 108

\[ u \]
Unity fingerprint 108

\[ v \]
valid molecular fields, generation of 199.
See also pharmacological similarity
– ACC approximation 199
– Coulombic interactions 199
– MM force 199
– molecular mechanics 199
– validating position and pattern of field points, tools for 200
– Cambridge Crystallographic Database 200
– XED model 199
Vanderbilt GlyT1 inhibitors 254
vBROOD system 180
topological structure, autocorrelation 119
virtual chemistry 83
virtual compounds, iterative generation 88
– analog generators 89
– chemical modifications, manual selection of 88, 89
– creation, different ways to 92
– inverse QSAR 89, 90
– multiple objective optimization 90
– structure-based de novo design 90–92
– transformations 88
virtual libraries 88, 92
– perspectives 98
virtual synthesis 92
– construction 94, 95
– iterative approaches 96
– outside big pharma 96
– readily synthesizable compounds 94
– real-life reactions 93
– searching 95
– synthetic tractability 92, 93
visualizations, of scaffold space 96, 97

\[ w \]
WOMBAT database 144
World Drug Alerts 242
XED force field 195, 200–202
– applications 200
– applied to scaffold hopping in Spark 202
– to benzene 203
– constructs
-- polarizable, and allowing for nuclear charge 201
-- types 200, 201
– creating new molecules 208–210
– desymmetrization of the ether oxygen atoms by 203
– electrons distribution 200

– far-reaching consequences of using molecular fields
-- as measures of similarity 212, 213
– improvement in electrostatic behavior 202, 204
– model reflecting intuitive molecular properties 199
– new potential inhibitors 209–212
– XED MM method 200
X-ray crystallography 279

Z
ZINC database 111, 152