UNIVERSITY OF CALIFORNIA
RIVERSIDE

Small Molecule Interaction With Biological Targets

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Bioengineering

by

Tyler William H Backman

December 2016

Dissertation Committee:

Dr. Thomas Girke, Chairperson
Dr. Jiayu Liao
Dr. Dimitrios Morikis

The Dissertation of Tyler William H Backman is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor Thomas Girke, for his mentorship and advice. I also thank Jiayu Liao, Dimitrios Morikis, and Victor G. J. Rodgers for support, advice, and constructive feedback. I thank Yiqun “Eddie” Cao for helping me get started in the field of cheminformatics. I thank Ronly Schlenk for extensively testing the prototype bioassayR database. I am greatly indebted to the many people whose excellent work I cite and build upon, and whose open source software tools and mathematical methods made this work possible. I also thank Samantha Lewis and Thomas Backman for inspiring and supporting me.

I acknowledge the support of the compute facility at the Institute for Integrative Genome Biology (IIGB) at UC Riverside. This work was supported by grants from the National Science Foundation [ABI-0957099] and the National Institute of Health [U24AG051129]. It was also supported by graduate fellowships from the University of California Office of the President (UCOP), the UC Riverside Graduate Division, and the National Science Foundation.

I acknowledge that the Oxford University Press (Nucleic Acids Research journal), and the American Chemical Society (Journal of Chemical Information and Modeling) granted me permission to use my work published in these journals (chapters 2 and 3 respectively) in my dissertation. I also acknowledge the Endocrine Society (Endocrinology journal), Elsevier (Neurobiology of Aging journal), and Taylor & Francis (Journal of Drug Targeting) for granting permission to reuse several figures, as mentioned in the figure legends, in this dissertation.

Thomas Girke is a coauthor on chapters 2, 3, 4, 5, and 6. Yiqun “Eddie” Cao is a coauthor of chapter 2.

You never change things by fighting the existing reality. To change something, design a new model that makes the existing model obsolete.
Richard Buckminster Fuller

ABSTRACT OF THE DISSERTATION

Small Molecule Interaction With Biological Targets

by

Tyler William H Backman

Doctor of Philosophy, Graduate Program in Bioengineering
University of California, Riverside, December 2016
Dr. Thomas Girke, Chairperson

In this dissertation, I present several strategies to leverage experimental data towards a quantitative understanding of small molecule bioactivity that can inform the discovery of small molecule drugs.

First, I present ChemMine Tools, a web service which provides both programmable and interactive online interfaces to a diverse set of analysis tools useful for analyzing small molecule structural data. ChemMine Tools allows users to import small molecule structures, compute pairwise compound similarities, search for similar compounds, cluster compounds by structure or physical properties, and compute physicochemical properties.

Second, I present bioassayR, a software package for large scale cross-target analysis of small molecule bioactivity profiles. bioassayR systematically analyzes data from thousands of screening experiments to identify target selective drug candidates and druggable protein targets. By simultaneously leveraging data from both custom small molecule screening efforts and public databases, bioassayR helps identify regions of the genome and proteome accessible to small molecule probes, elucidate novel mechanisms of action for bioactive molecules, and predict off-target effects which currently lead to a high attrition rate in drug discovery efforts.

Third, I present a systematic analysis of small molecule target selectivity against druggable protein targets in large public bioactivity data. This study highlights a large number of novel bioactivity patterns with the potential to inform the curation of better drug discovery libraries. I demonstrate that a large fraction of the previously-reported promiscuity of FDA approved drugs is due to cross-reactivity among protein targets sharing common domains and/or Molecular Function annotations. I also identify a large number of novel active targets for the FDA approved drugs, as well as a large set of novel compounds active against potentially new therapeutic targets with no evidence of druggability by FDA approved drugs.

Finally, I present a mathematical model of leptin transcytosis across the blood brain barrier. This kinetic model extends current mathematical models of receptor endocytosis to transcytosis, and behaves similar to the experimentally observed dynamics of this system. A computational model is provided which allows for *in-silico* perturbation, to predict the potential effects of pathological states, or therapeutic small molecules.

Contents

List of Figures **xii**

List of Tables **xviii**

1 Introduction **1**

1.1 Bibliography 8

2 ChemMine Tools: an Online Service for Analyzing and Clustering Small Molecules **10**

2.1 Abstract 10

2.2 Introduction 11

2.3 Methods 14

2.4 Discussion of Services 15

2.4.1 *Compound Workbench* 15

2.4.2 *Similarity Toolbox* 16

2.4.3 *Search Toolbox* 17

2.4.4 *Clustering Toolbox* 17

2.4.5 *Property Toolbox* 19

2.5 Conclusion and Further Development 20

2.6 Bibliography 22

3 bioassayR: Cross-Target Analysis of Small Molecule Bioactivity **26**

3.1 Abstract 26

3.2 Introduction 27

3.3 Methods and Implementation 29

3.3.1 Software design and workflow overview 29

3.3.2 bioassay object: importing data 30

3.3.3 bioassayDB object: multiple assay SQL database 32

3.3.4 bioassaySet object: storing multiple assays in a matrix 33

3.3.5 Pre-built PubChem BioAssay database 34

3.3.6 Identifying compounds with selective or promiscuous bioactivity 36

3.3.7 Clustering small molecules by bioactivity profile 37
3.4 Results and Discussion 39
3.4.1 Abbreviations 41
3.5 Bibliography 42

4 bioassayR: Supporting Information 47
4.1 Use case examples 48
4.1.1 Drug-target space analysis 49
4.1.2 Clustering drugs by structure and bioactivity profile 51
4.1.3 Guiding high-content screens with reference data from single target bioassays 52
4.2 Time performance evaluation 56
4.3 Bibliography 58

5 Large-scale Bioactivity Analysis of the Druggable Proteome 60
5.1 Abstract 60
5.2 Introduction 61
5.3 Results and Discussion 66
5.3.1 Bioactivity Data 66
5.3.2 Target Selectivity 76
5.3.3 Promiscuous Binders and Hit Ratio Statistical Model 86
5.3.4 Comparison Between Annotated Drug Targets and Public HTS Data 94
5.3.5 Drug-Target (DT) Biclustering Analysis 96
5.3.6 Target-Protein (TP) Network 104
5.4 Methods 106
5.4.1 Bioactivity Database 106
5.4.2 Clustering Protein Targets by Sequence 106
5.4.3 Clustering of Compound-Target Matrix 107
5.4.4 Protein Annotations and GO Enrichment 107
5.4.5 BicBin Biclustering Of Drug-Target Matrix 108
5.4.6 Hit Ratio Bayesian Model and Mixture Distribution 109
5.4.7 Promiscuous Aggregators 111
5.4.8 Pan-Assay Interference Compounds (PAINS) 111
5.4.9 Target-Protein (TP) Network and Network Visualizations . . . 112
5.5 Conclusion 113
5.6 Supporting Information 116
5.7 Bibliography 119

6 Supporting Information for Large-scale Bioactivity Analysis of the Druggable Proteome 124
6.1 Target Selectivity Distribution 124

6.2 Fully Screened Sub-Matrix 125
6.2.1 Fully Screened Sub-Matrix Methods 127
6.3 Error Rate Estimate 127
6.4 Target Selectivity by Molecular Size 130
6.5 Stretched Exponential Selectivity Distribution 133
6.5.1 Stretched Exponential Selectivity Distribution Methods 136
6.6 Additional Figures 136
6.7 Bibliography 140

7 Receptor-Mediated Leptin Transcytosis: a Model of the Blood-Brain Barrier as a Lipostatic Regulatory Interface 142
7.1 Background and Introduction 142
7.1.1 Transport Physiology 145
7.2 Methods (Transport Kinetics) 148
7.2.1 Quasi-steady-state Transport Rate 148
7.2.2 Diffusion vs Transport Rate 150
7.2.3 Dynamic Behavior and Receptor Regulation 153
7.2.4 Computational Model 154
7.3 Results and Discussion 158
7.3.1 Quasi-steady-state Behavior 158
7.3.2 Dynamic Behavior 160
7.4 Further Work 161
7.5 Bibliography 164

8 Conclusion and Future Vision 167
8.1 Overview 167
8.2 ChemMine Tools 168
8.3 bioassayR 170
8.4 Large-scale Bioactivity Analysis of the Druggable Proteome 172
8.5 Leptin Transcytosis 175
8.6 Future Vision 176
8.7 Bibliography 180

List of Figures

2.1 **Illustration of the functionalities provided by *ChemMine Tools*.** The utilities of the five application domains (I-V) are listed in more detail in Table 2.1. 13

3.1 **Design overview and workflow.** bioassayR stores bioactivity data in four interconnected objects. **A** Data from a single bioassay experiment is imported into a bioassay object. **B** Any number of bioassay objects can be loaded into the bioassayDB SQL database which is optimized for time efficient searching. **C** Filter and query methods are available to identify compounds or assays of interest. These query results can be imported into a bioassaySet object which stores activity data as a sparse matrix where columns represent compounds and rows assays (targets). This organization facilitates many typical cross-target analysis routines, *e.g.* target selectivity analyses. **D** To reduce both redundancy and sparseness in the data, assays involving the same or similar targets can be collapsed into a single row using the *perTargetMatrix* function. 31

4.1 **Frequency of active PubChem BioAssay compounds across GO slim categories.** The target proteins represented in PubChem BioAssay have been classified by GO slim terms of the Molecular Function Ontology (vertical axis). The number of compounds with activity or inactivity against at least one target protein within each protein class is plotted on the horizontal axis. The plot on the left gives the numbers for FDA approved drugs, the one in the center for all other compounds, and the one on the right gives the numbers of the corresponding protein targets included in the assays. To distinguish between small and zero quantities of active compounds, the active point is omitted for GO slim terms which have no active compounds. 50

4.2 **Clustering small molecules by structure and bioactivity profile.** Multidimensional scaling (MDS) was used to embed small molecules into a two dimensional space (x- and y-axis). Each point represents an FDA approved drug. The distance between the points is proportional to the similarity of the bioactivity profile available for each drug. Structure-based clustering results are also indicated by colors for 8 mid-size clusters, while the remaining compounds are shown in grey. The latter clustering used atom pair fingerprints as similarity measure and hierarchical clustering with tree cutting to assign compounds to discrete structure-based similarity groups. Four example compounds are shown, which are members of the same structural cluster (shown here in purple), which is comprised largely of steroid drugs. 55

5.1 **Bioactivity data mining strategy.** Public bioactivity data was first summarized in a compound-target bioactivity matrix (**A**). Protein targets and small molecules were clustered by sequence (**B**) and structure (**C**) respectively, and compound-target sets with shared bioactivity profiles were identified with biclustering (**D**). For small molecules, the distributions of (**E**) target selectivity (the number of active targets) and (**F**) hit ratio (the fraction of screened targets that are active) were quantified. For protein targets, enriched GO (Gene Ontology) terms (**G**) among proteins with common bioactivity were identified, and a network (**H**) was constructed which connects target proteins with similar bioactivity profiles. These analyses highlight several interesting bioactivity patterns, identify promiscuous and selective compounds, and identify druggable protein targets and protein domains. 64

5.2 **Pfam domain screening participation.** The quantity of Pfam domains is plotted on the horizontal (x) axis whose protein targets have at least as many active or inactive compounds as shown on the vertical (y) axis. The left panel includes all Pfam domains in the PubChem BioAssay targets, while the right panel includes just those domains also present in the *H. sapiens* proteome, including non-*H. sapiens* targets which share a common domain with an *H. sapiens* protein. 73

5.3 **Frequency of active PubChem BioAssay compounds across protein target domains.** The target proteins represented in PubChem BioAssay have been classified by Pfam protein domains present in the *H. sapiens* proteome (vertical axis). We report data for all proteins which encode a Pfam domain present in the *H. sapiens* proteome, even if the assay was performed against a protein from another species. We show here only domains with at least 100 amino acid residues in the homology model, to avoid small repeats and domains unlikely to be drug targets. The quantity of targets with each domain among the PubChem BioAssay data, and within the *H. sapiens* proteome (all proteins, including those not screened in PubChem BioAssay) are shown on the right in both plots. **(A)** The top 35 Pfam domains with the greatest number of active FDA approved drugs, in decreasing order. **(B)** The top 35 Pfam domains with the greatest number of non-FDA compounds, but no active FDA approved drugs, in decreasing order. A full table with the number of active compounds for each domain, including non-*H. sapiens* domains, and domains with under 100 residues is provided in the S2 File of Supporting Information. 75

5.4 **Distribution of active proteins per compound.** Both plots show the distribution of target selectivity among PubChem BioAssay compounds, with selectivity quantified by three methods which cluster protein targets across increasingly large evolutionary distances, as described in the text. FDA approved drugs (red) are shown separately from non-FDA compounds (blue). **(A)** Semi-log plot of the target selectivity distributions, where horizontal (x) axis represents the number of active protein targets and/or protein target clusters, while the vertical (y) axis represents the fraction of each compound set that is active against a particular number of targets. **(B)** Box plot of the target selectivity distributions, with horizontal lines at the 25%, 50%, and 75% quantiles for each distribution. The vertical (y) axis represents the number of active protein targets and/or protein target clusters. Whiskers extend to 1.5 times the inter-quartile range, however we limit the vertical (y) axis to 16 in order to zoom into the higher density region. 78

5.5 **Molecular Function Gene Ontology (MF GO) slim Term Enrichment vs Domain Selectivity.** Pfam domains are binned by the median domain selectivity of active compounds against targets with these domains, as in Table 5.3. The domains in each bin were computed separately based on FDA approved and non-FDA compounds, shown here side by side. For each bin of domain selectivity, the enrichment of MF GO slim terms against the background of all bins is shown. Enriched terms are sorted increasingly by the lowest p-value obtained, with all terms shown here having a p-value < 0.05. The right column dot plot shows the number of protein targets in PubChem BioAssay annotated with each MF GO slim term. 85

5.6 **probability density of hit ratios (θ)** Shown here is an equally weighted convex combination of hit ratio probabilities for individual compounds, which represents the probability of any individual compound from a set having a specific hit ratio. Smoothing was applied to reduce sampling noise in low probability regions. **(A)** Hit ratio distributions for FDA approved compounds vs non-FDA. **(B)** Hit ratio distributions for aggregator compounds vs non-aggregators. **(C)** Hit ratio distributions for PAINS vs non-PAINS. 89

5.7 **Bioactivity of drug-target bicluster 1.** The vertical axis lists the drugs in this bicluster by common name, and the horizontal axis represents the UniProt names for the representative targets of each sequence-similar target cluster. The compound-target pairs are colored according to one of six colors: untested in PubChem BioAssay (black), inactive in PubChem BioAssay (grey), active in PubChem BioAssay (dark green), untested but annotated as active in DrugBank (green), inactive in PubChem BioAssay but annotated as active in DrugBank (blue), and active and also annotated as active in DrugBank (light green). Rows and columns are sorted by bioactivity profile similarity. 99

5.8 **Molecular Function Gene Ontology Slim (MF GO Slim) term enrichment for each drug-target bicluster.** Enrichment measured by hypergeometric test. Terms with $p \leq 0.05$ are shown and sorted increasingly. 101

5.9 **Bicluster (color) vs compound structure (position).** Multidimensional scaling (MDS) was used to embed small molecules into a two dimensional space (x- and y-axis). Each point represents an FDA approved drug. A density map colored by each bicluster is shown for the MDS principal coordinate 1 (on top) and principal coordinate 2 (on right). The distance between the points is proportional to the chemical similarity between the two compounds. Bioactivity-based biclustering results are also indicated by colors, with each compound assigned to it's lowest numbered (densest and/or largest) bicluster. Only the 11 biclusters with the largest number of compounds are shown, to allow for a visually distinct color palette. 103

5.10 **Target-protein network and Molecular Function Gene Ontology Slim (MF GO Slim).** Each node represents a protein target, and edges connect any two protein targets with greater than 50% bioactivity similarity across non-promiscuous binding compounds. Targets are colored according to MF GO Slim terms, with unannotated targets colored black. Shown are 2407 nodes (target proteins with at least one edge) and 11317 edges representing shared bioactivity among the mutually screened subset of the 1.2M compounds tested in the bioassays we analyze in this study. 105

6.1 **Distribution of distinct protein target assay participation.** Data is included from all assay experiments in PubChem BioAssay annotated with one or more clearly defined protein targets, and reporting an active score for at least one small molecule. The dashed vertical line is drawn at 10 targets, which is the minimum value we categorize in this study as a “highly screened” compound. 126

6.2 **Target selectivity by molecular size.** Violin plot with horizontal lines drawn at the 0.25, 0.5, 0.75 quantiles with tails trimmed to the range of data. Molecule size is quantified here by the number of non-hydrogen (heavy) atoms. (A) Target selectivity vs. molecular size across the entire range (y axis) of variation in these data. (B) Target selectivity vs. molecular size zoomed in on the y-axis to show more detail. 132

6.3 **Selectivity function fit.** The distribution of distinct sequence targets for non-FDA approved compounds, along with best fit lines using two-parameter versions of the exponential, power law, and stretched exponential functions. 134

6.4 **Drug-Target (DT) bipartite network biclusters.** Protein targets are shown in black, with FDA approved drugs shown in color, based on their bioactivity bicluster. Unclustered compounds are shown in grey. No color key is provided, as some colors were reused in order to visualize a large number of biclusters. Node position is based on connectivity, with the same positions as in Fig 6.5. 137

6.5 **Drug-Target (DT) bipartite network Gene Ontology (GO).** FDA approved drugs are shown in black, with protein targets show in color based on the most specific Molecular Function GO Slim term for each target. Unannotated targets are shown in white. No color key is provided, as some colors were reused in order to visualize a large number of GO terms. Node position is based on connectivity, with the same positions as in Fig 6.4. 138

6.6 **Sensitivity of PAINS and aggregators vs promiscuity probability cutoff.** The top panel shows the sensitivity (true positive rate) of PAINS and aggregators to categorize promiscuous compounds throughout a range of promiscuity probability cutoffs $P(\theta \geq 0.25) > x$ over the range $x = [0.01, 0.9999]$. The bottom panel shows the number of promiscuous compounds at each cutoff value. 139

7.1 **Illustration of a generalized clarathrin-dependent receptor mediated transcytosis mechanism.** Figure reprinted from Smith and Gumbleton 2006 with permission from the publisher[21]. 146

7.2 **Equation** (7.6)**: Unit-less saturable response of transport rate to serum leptin levels.** 151

7.3 **Diagram of diffusion vs transport in the rat brain capillary.** v_{max} represents the maximal rate of transport, D the binary diffusion coefficient of leptin in blood, C the serum leptin concentration, and L the capillary radius. Figure adapted from Hicks et al. 1983 with permission from Elsevier[8]. 152

7.4 **An example fit of the leppRedict computational model to artificial data over a one hour time course beginning with all initial leptin bound to surface receptors.** Dots represent artificial data with random error applied, and solid lines show the model behavior as fit to these data. This figure illustrates the ability of the model to fit and visualize arbitrary data (and eventually experimental data); not a prediction of the system behavior. 157

7.5 **Saturable response of transport rate to serum leptin levels at six sites from *in vivo* rat models.** Sites of slow uptake are unlikely to contribute significantly to leptin transport under physiological conditions. Figure republished from Zlokovic et al. 2000 with permission from the Endocrine Society[32]. 159

List of Tables

2.1 **List of services provided by *ChemMine Tools.*** The names of software tools, libraries and environments are italicized. Programs developed by the *ChemMine Tools* project are labeled with a star*. Acronyms defined in text. 13

4.1 **Small molecule databases.** Here we provide links to several small molecule structure and/or bioactivity databases that can potentially be used with bioassayR. This list is not comprehensive. 47

4.2 **Selected cross-target analysis functions.** Instructions and examples are provided in the online package documentation available at http://bioconductor.org/packages/bioassayR/. 48

4.3 **Results of a high-content screen compared to single target bioassays of the same pathway.** [1]The numbers of total and active compounds are given for the high-content PubChem BioAssay #465. This assay screened for inhibitors of antigen receptor-induced NF-kappaB activation. [2]The same numbers are given for the subset of compounds that have also been found active in single target assays where the corresponding target protein is associated with NF-kappaB processes. Currently, 605 single target assays meet this requirement. The results are subdivided into the number of compounds active in variable numbers of single target assays including $\geq$1 and 1-5. 54

4.4 **BioassayR function runtime.** Total runtime in seconds for seven key bioassayR functions over the range of 1-10,000 randomly selected compounds. Results generated on a single cpu core of an AMD Opteron 6376 2.3 Ghz processor. 56

5.1 **Screening frequency of FDA approved and non-FDA compounds against increasing numbers of protein targets.** Data is included from all assay experiments in PubChem BioAssay annotated with one clearly defined protein target, and reporting an active score for at least one small molecule. Multiple assays against the same target are counted only once. 68

5.2 **Screening Frequency.** The number of distinct compound-target combinations screened in multiple assays, listed for increasing numbers of assays. 70

5.3 **Frequency of Pfam domains binned by median domain selectivity of active compounds.** Each row represents a set of Pfam domains whose active compounds (against targets with that domain) have a median domain selectivity in the range specified. Domain selectivity is the same as introduced in the "Target Selectivity Distribution" section above, where active targets sharing a common domain are counted only once. The ranges are ordered from top to bottom by increasing number of distinct domain active targets. We report bin counts separately for FDA Approved and Non-FDA compounds. . . . 81

5.4 **Top 15 Pfam domains with least selective active drugs.** Only domains present in the *H. sapiens* proteome are shown, and are sorted and selected by decreasing domain selectivity among FDA approved drugs. We also excluded domains with under 100 amino acid residues in the homology model, to avoid small repeats and domains unlikely to be drug targets. To the left of the slash in each column is the median target selectivity of FDA approved compounds active against this domain, and to the right is the median selectivity of non-FDA compounds. Selectivity is quantified by three methods of increasingly grouped protein targets as described in the text. For example, the non-FDA approved compounds active against targets with the PF00183 domain are active against a median of 12 targets with distinct GenBank identifiers, so this domain has a 12 after the slash in the first column. 82

5.5 **Top 14 Pfam domains with most selective active drugs.** Only domains present in the *H. sapiens* proteome are shown, and are sorted and selected by increasing domain selectivity among FDA approved drugs. We also excluded domains with under 100 amino acid residues in the homology model, to avoid small repeats and domains unlikely to be drug targets. To the left of the slash in each column is the median target selectivity of FDA approved compounds active against this domain, and to the right is the median selectivity of non-FDA compounds. Selectivity is quantified by three methods of increasingly grouped protein targets as described in the text. For example, the non-FDA approved compounds active against targets with the PF00144 domain are active against a median of 6 targets with distinct GenBank identifiers, so this domain has a 6 after the slash in the first column. 83

5.6 **Comparison of PubChem BioAssay activity data to DrugBank target annotations.** All compound-target pairs for FDA Approved drugs are grouped into one of six possible categories. Depending on the HTS results in PubChem BioAssay, a compound-target pair is annotated as either untested, inactive, or active (rows in this table). Additionally, the compound-target pair is either annotated or unannotated as a known active target in DrugBank (columns in this table). Counts outside of parenthesis represent results against all protein targets, whereas counts to the right in parenthesis represent results against the subset in which the representative UniProt indentifer for each target cluster is from the *H. sapiens* proteome. 96

5.7 **Top Pfam domains in each bicluster.** Shown are the top 16 highest scoring drug-target biclusters with more than one compound and more than one target. The number of drugs (cids) and targets is shown in columns 2 and 3, respectively. The 4th and 5th columns give the name of the most abundant domain and its frequency, respectively. The last (6th) column shows the BicBin score, representing the density and size of the bicluster. The BicBin score is the negative exponent of the Chernoff Bound. It is inversely proportional to the probability of each bicluster occurring by random chance, as described in Methods. . . . 98

6.1 **Distribution of active proteins per compound.** Selectivity is quantified by three methods which cluster protein targets across increasingly large evolutionary distances, as described in the text. Mean values include a standard deviation after the ± symbol. Trimmed mean represents the mean of compounds with values of 20 or fewer. We include a trimmed mean, as mean selectivities are highly skewed by a small number of highly promiscuous compounds, whose selectivity counts reflect the screening data volume rather than the bioactivity profiles. 124

6.2 **Disagreement and agreement among compound target pairs screened 2-4 distinct times.** Shown are the number of pairs that had a given number of inactive (vs active) results in these replicates. For example if a given protein target pair was screened in two assays (Times Screened = 2) and both results were inactive it's count would be added to the "2 Inactives" column. If the two assays disagree and one showed active and the other inactive, it is counted in the "1 Inactive" column. 128

6.3 **Variation in molecular size for FDA approved and non-FDA compounds.** FDA approved compounds have a slightly lower mean size than non-FDA compounds. Only highly screened active compounds were included in this calculation. 131

6.4 **Target selectivity of the 40 largest highly screened active FDA Approved Drugs.** Sorted by decreasing size (non-hydrogen atom count). The description column represents one common clinical use of each compound, but may not represent its only known therapeutic utility. 133

7.1 **Saturable response of transport rate to serum leptin levels at six sites from *in vivo* rat models.** Choroid Plexus data represents binding to tissue rather than direct transport. Data from Zlokovic et al. 2000 [32]. 160

Chapter 1

Introduction

This work presents three substantially different approaches towards the common goal of quantitatively leveraging experimental data and biological understanding towards the discovery and development of safe and effective medical therapies. First, I present new software tools for the structural analysis of drug-like small molecules. Second, I present new software tools for the large scale cross-target analysis of experimentally obtained small molecule bioactivity data, along with an extensive analysis of the bioactivity data that is publicly available. Lastly, I present a case study of translating a qualitative understanding of a disease related biological process into a quantitative model which can be computationally interrogated to identify druggable nodes in a complex biological system, and to identify gaps in our knowledge that make our current models inadequate to explain experimental observations. While independently useful, the ultimate utility of these three methods will be realized as they are further developed into a unified workflow.

I first present ChemMine Tools, a computational environment for visualizing, searching, clustering, and predicting physicochemical properties of small molecules[3]. This work provides both a web interface, and a programmable R language interface embedded into the ChemMineR cheminformatics library[5]. There is also a standalone version that users can run on a desktop computer or private server. A key feature of ChemMine Tools is it's modularity, as it provides a common API for developers to easily contribute additional tools, which simultaneously become available to web and R users. Users on both interfaces submit jobs to a queue, which runs a large number of parallel jobs in the background, while maintaining a responsive interface. After a job is completed, the results are returned in both a downloadable machine-readable form, and in interactive plots and tables. Since the release of the current version, ChemMine Tools has become a widely used foundational tool in drug discovery, and has been cited in the discovery of a large number of small molecule drugs and chemical genomics probes.

Second, I present bioassayR, a computational tool for large scale data mining of small molecule bioactivity data from diverse sources. Despite a large and rapidly growing body of small molecule bioactivity data, systematic leverage of these data as a reference for identifying compounds with a desired bioactivity, and assessing the druggability of protein targets are limited by informatics challenges stemming from the large data volume, heterogenous experimental designs, sparseness, and noise. bioassayR addresses these issues to enable simultaneous analysis of thousands of bioassay experiments performed over a diverse and sparse set of compounds and biological targets. bioassayR works with both public and user supplied data and is freely available as an open-source R/Bioconductor package.

Organizing and systematically analyzing large bioactivity data presents a unique set of technical challenges. As public bioactivity databases include hundreds of millions of compound-target activity outcomes, the data is too large to analyze efficiently with standard data structures provided by existing software environments and instead is better suited to a custom-engineered relational database, and/or sparse matrices. Sparseness due to untested compound/target combinations presents challenges in searching and clustering small molecules by their activity profile, as most such algorithms are designed for working with a complete matrix. Structural similarity amongst compounds are often represented by metrics such as the Tanimoto Coefficient quantifying similar and distinct entries in a binary matrix representing the presence or absence of specific features[6]. Comparing similarity of bioactivity profiles between compounds requires statistical methods capable of handling data sparseness as many molecule pairs have not been screened against a uniform set of targets. Signal to noise ratio is also a major consideration, as the false positive and false negative rates for bioassays are notoriously high[7]. bioassayR implements a large number of new data mining strategies to address these challenges.

By leveraging the capabilities of bioassayR, I presents the results of a large-scale data mining project analyzing small molecule bioactivity patterns across millions of compound vs protein target activity outcomes in the PubChem BioAssay database[19]. The goal of this analysis is to identify patterns of small molecule target selectivity and bioactivity that can inform the costly drug discovery and development process, which has a high attrition and recall rate[1, 16, 2, 15]. Historically, most new pharmaceuticals have been discovered based on an observed phenotype *in vivo*, however substantial advances have been made in target-based discovery, where small molecule drugs are chosen based on patterns of bioactivity against protein targets which influence disease-related processes. Target-based “rational” drug discovery has led to the discovery of a substantial number of new pharmaceuticals, however the effectiveness and throughput of this process continues to lag behind *in vivo* phenotype driven discovery, in producing safe and effective drugs that ultimately obtain FDA approval[18, 17].

Many current FDA approved drugs were discovered by chance, or by screening live organisms for a desired phenotype, and their exact mechanism of action remain unknown[18]. A small subset of existing drugs were discovered using target based strategies, where a desired molecular mechanism and target protein(s) are first identified, and experiments are performed to identify molecules with the desired activity. Target based drug discovery approaches are severely limited by the lack of detailed knowledge regarding the mechanisms of existing drugs. For example- do existing drugs tend to bind to a single molecular target, multiple related targets, or many different targets? This question is extensively explored in Chapter 5.

PubChem Bioassay was started in 2004 as part of the NIH’s Roadmap for Medical Research Initiative, and is currently the largest repository of public domain bioactivity data. In recent years, it has been expanded to also include the contents of BindingDB and ChEMBL[20, 9, 13]. As of July 2016, PubChem Bioassay contains the results of over 1.2 million bioassay experiments. About 70000 of these are defined target assays screening approximately 1 million distinct compounds against approximately 6k distinct protein targets. These defined assay data are a useful reference for determining patterns of small molecule binding activity, as a significant portion (over 400k compounds) have been screened against at least 10 distinct targets, and show activity against at least one protein target.

I build in part on the discoveries of three key PubChem BioAssay data mining projects, which investigate the patterns of small molecule bioactivity across many protein targets in large screening data. In 2009 Han *et al.* reported the distribution of assay participation, and target selectivity in this data, and also found that over 50% of compounds are active only against sequence similar targets[10]. In 2013 Hu *et al.* reported that most active compounds from biological screening have lower promiscuity than FDA approved drugs[11]. Recently, Jasial *et al.* found many surprising new patterns, including the observation that most non-FDA approved bioactive compounds from previous screens continue exhibiting low promiscuity as they become more extensively assayed over time[12].

I provide additional context to the discoveries mentioned above, by analyzing a larger set of data with new informatics tools, and identifying several patterns of bioactivity with a potential to improve the curation of drug discovery compound libraries. I find a large number of novel active targets for the FDA approved drugs, as well as a large set of novel compounds active against potentially new therapeutic targets with no evidence of druggability by FDA approved drugs. By quantifying the rate of agreement between replicated compound-target activity outcomes, I estimate that less than half of these novel outcomes are due to experimental and data curation errors. Using a statistical model of promiscuity, I demonstrate that a large fraction of the previously reported high average level of promiscuity of FDA approved drugs is due to cross-reactivity between related protein targets, and is not due to biased screening volume. I show that a large fraction of these multi-target drugs exhibit cross-reactivity within biclusters, where a common set of compounds are active against a common set of targets sharing related domains and molecular function annotations. I also identify the set of highly screened compounds with strong experimental evidence for promiscuous binding across many targets, and compare and contrast several common methods of identifying promiscuous compounds, which often lead to problematic false positives in drug discovery efforts. Aggregator assays identify a small number of highly promiscuous compounds, while PAINS substructures identify a larger set of somewhat less promiscuous compounds[8, 14, 4]. I also report a large number of compounds with strong experimental evidence of promiscuous binding that were not detected by either method.

The systematic analysis of public domain bioactivity data in this dissertation provides insight into many outstanding questions regarding drug discovery and bioactivity. Simultaneously, the software tools I developed in the process address the computational challenges mentioned above, in a manner that will enable other researchers to efficiently answer questions about patterns of small molecule target activity, and to narrow the search space when identifying drug and small molecule probe candidates. Organized information on the active targets of known drugs and drug like small molecules, as well as the functional relationships between these active targets will enable the identification of target selective small molecules, as well as multi-target drug candidates with potential to influence robust pathways and cellular networks. Additionally, synergistic drug combinations may be predicted by combinatorial use of highly target selective small molecules which together exhibit a desired activity pattern.

Lastly, I present a mathematical model of transcytosis of the peptide hormone leptin across the blood brain barrier. This serves as an example of taking a well developed qualitative understanding of a biological process, and translating that understanding into a system of differential equations that can be simulated computationally. I show that this model can be fit to appropriate experimental data to infer specific values of the underlying model parameters necessary to make predictions about the system behavior under dynamic conditions. A similar method has been demonstrated to successfully measure the unknown rate constants in receptor-mediated cellular uptake of epidermal growth factor (EGF) by it's receptor (EGFR)[21]. This model provides several key scientific opportunities. First, a sensitivity analysis can systematically identify which currently unknown aspects of the underlying system must be further studied to enable meaningful predictions of the system behavior under diverse conditions, and which can be safely ignored or approximated. Additionally, the same process can identify nodes in the system (such as enzymatic proteins) that exhibit potentially druggable sensitivity to perturbation, and other system nodes which exhibit robustness to perturbation.

1.1 Bibliography

[1] David J Adams. The Valley of Death in anticancer drug development: a reassessment. *Trends in Pharmacological Sciences*, 33(4):173–180, April 2012.

[2] Ravi Iyengar Aislyn DW Boran. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3):297, May 2010.

[3] Tyler W H Backman, Yiqun Cao, and Thomas Girke. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic acids research*, 39(Web Server issue):W486–91, July 2011.

[4] Jonathan B Baell and Georgina A Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.*, 2010.

[5] Yiqun Cao, Anna Charisi, Li-Chang Cheng, Tao Jiang, and Thomas Girke. ChemmineR: a compound mining framework for R. *Bioinformatics (Oxford, England)*, 24(15):1733–1734, August 2008.

[6] Xin Chen and Charles H Reynolds. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of chemical information and computer sciences*, 42(6):1407–1414, November 2002.

[7] Thomas J Crisman, Christian N Parker, Jeremy L Jenkins, Josef Scheiber, Mathis Thoma, Zhao Bin Kang, Richard Kim, Andreas Bender, James H Nettles, John W Davies, and Meir Glick. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *Journal of chemical information and modeling*, 47(4):1319–1327, July 2007.

[8] Brian Y Feng, Anton Simeonov, Ajit Jadhav, Kerim Babaoglu, James Inglese, Brian K Shoichet, and Christopher P Austin. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.*, 2007.

[9] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(Database issue):D1100–7, January 2012.

[10] L Han, Y Wang, and S H Bryant. A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics*, 25(17):2251–2255, September 2009.

[11] Ye Hu and Jürgen Bajorath. High-resolution view of compound promiscuity. *F1000Res*, 2013.

[12] Swarit Jasial, Ye Hu, and Jürgen Bajorath. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS ONE*, 2016.

[13] T Liu, Y Lin, X Wen, R N Jorissen, and M K Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, 35(Database issue):198–201, January 2007.

[14] S L McGovern, B T Helfand, B Feng, and B K Shoichet. A specific mechanism of nonspecific inhibition. *J. Med. Chem.*, 46(20):4265–4272, September 2003.

[15] Bernard Munos. Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discovery*, 2009.

[16] Deendayal Dinakarpandian Sachin Mathur. Drug Repositioning using Disease Associated Biological Processes and Network Analysis of Drug Targets. *AMIA Annual Symposium Proceedings*, 2011:305, 2011.

[17] Frank Sams-Dodd. Target-based drug discovery: is something wrong? *Drug discovery today*, 10(2):139–147, January 2005.

[18] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature Reviews Drug Discovery*, 10(7):507–519, June 2011.

[19] Yanli Wang, Tugba Suzek, Jian Zhang, Jiyao Wang, Siqian He, Tiejun Cheng, Benjamin A Shoemaker, Asta Gindulyte, and Stephen H Bryant. PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, 2013.

[20] Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A Shoemaker, Evan Bolton, Asta Gindulyte, and Stephen H Bryant. PubChem's BioAssay Database. *Nucleic acids research*, 40(Database issue):D400–12, January 2012.

[21] H Steven Wiley, Stanislav Y Shvartsman, and Douglas A Lauffenburger. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol.*, 2003.

Chapter 2

ChemMine Tools: an Online Service for Analyzing and Clustering Small Molecules

2.1 Abstract

ChemMine Tools is an online service for small molecule data analysis. It provides a web interface to a set of cheminformatics and data mining tools that are useful for various analysis routines performed in chemical genomics and drug discovery. The service also offers programmable access options via the R library ChemmineR. The primary functionalities of ChemMine Tools fall into five major application areas: data visualization, structure comparisons, similarity searching, compound clustering and prediction of chemical properties. First, users can upload compound data sets to

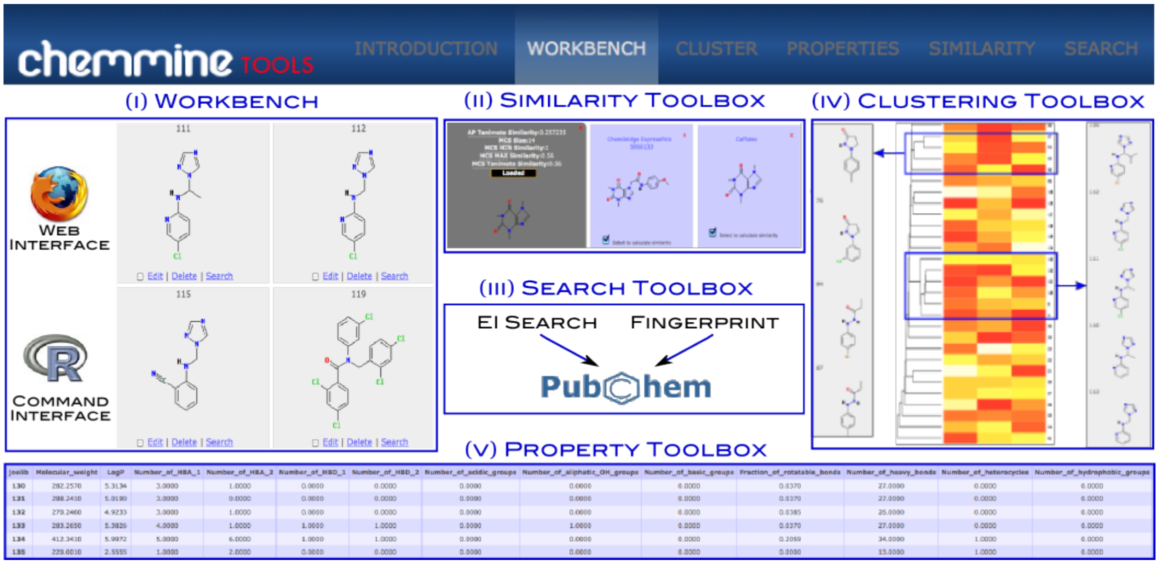
the online Compound Workbench. Numerous utilities are provided for compound viewing, structure drawing and format interconversion. Second, pairwise structural similarities among compounds can be quantified. Third, interfaces to ultra-fast structure similarity search algorithms are available to efficiently mine the chemical space in the public domain. These include fingerprint and embedding/indexing algorithms. Fourth, the service includes a Clustering Toolbox that integrates cheminformatic algorithms with data mining utilities to enable systematic structure and activity based analyses of custom compound sets. Fifth, physicochemical property descriptors of custom compound sets can be calculated. These descriptors are important for assessing the bioactivity profile of compounds in silico and QSAR analyses. ChemMine Tools is available at: `http://chemmine.ucr.edu`.

2.2 Introduction

Cheminformatics tools for analyzing small molecule screening data play an important role in many fields including chemical biology, chemical genomics, drug discovery and agrochemical research [37, 17, 31]. Informatics resources in these areas are essential for exploring the structure, properties and bioactivity of biologically relevant molecules. To provide these capabilities, software tools are required for analyzing the structural similarities, physicochemical properties and bioactivity profiles of natural and synthetic compounds to gain insight into their modes of action in biological systems. This information is important for the development of effective small molecule probes for studying the functions of protein and cellular networks in chemical genomics and drug discovery research [11]. In addition, similar informatics resources are required for identifying the structural and physicochemical relationships among compounds from metabolic or signaling pathways [19, 23, 32]. The rapidly growing relevance of chemical genomics approaches for modern biology research has significantly increased demand for small molecule mining systems in academia [30].

Currently, the structures of over 30 million distinct small molecules are available in open-access databases, including *PubChem*, *ChemBank* and many others [1, 34, 21, 13, 7, 22, 25]. In addition, preliminary bioactivity data from hundreds of high-throughput screening (HTS) experiments against a wide spectrum of target sites have become available for almost one million compounds in the bioassay sections of various public databases (see below). To efficiently analyze these resources, the development of novel compound data mining and cheminformatic web services is essential.

While there has been extensive development of public domain small molecule databases in recent years [40, 21, 9, 1, 25, 34, 23, 27, 42, 10, 3, 24, 44, 7, 22, 14], the number of open access web services for analyzing public or custom small molecule data is extremely limited at this point [35, 45]. Thus far, most development has been focused on standalone software applications targeted toward computational rather than experimental scientists. These include *Open Babel* [16, 29], the *Chemistry Development Kit* [36, 15], the *Chemical Descriptors Library* [38] and *JOELib* [43]. Examples of software designed for non-expert users in this field are Chembench [41] for online QSAR modeling and KNIME [2] for designing data analysis pipelines.



joelib	Molecular_weight	LogP	Number_of_HBA_1	Number_of_HBA_2	Number_of_HBD_1	Number_of_HBD_2	Number_of_acidic_groups	Number_of_aliphatic_OH_groups	Number_of_basic_groups	Fraction_of_rotatable_bonds	Number_of_heavy_bonds	Number_of_heterocycles	Number_of_hydrophobic_groups
130	282.2570	5.3134	3.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0370	27.0000	0.0000	0.0000
131	288.2410	5.0190	3.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0370	27.0000	0.0000	0.0000
132	270.2460	4.9233	3.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	26.0000	0.0000	0.0000
133	283.2650	5.3826	4.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0370	27.0000	0.0000	0.0000
134	412.3410	5.9972	5.0000	6.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.2059	34.0000	1.0000	0.0000
135	220.0010	2.5555	1.0000	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	13.0000	1.0000	0.0000

Figure 2.1: **Illustration of the functionalities provided by *ChemMine Tools*.** The utilities of the five application domains (I-V) are listed in more detail in Table 2.1.

Functions	Program	Input	Output	Comments
(I) Compound Workbench				
Structure import/export	*Open Babel*	mouse clicks	SMILES/SDF	one or many compounds
Format interconversions	*Open Babel*	SDF/SMILES	SMILES/SDF	one or many compounds
Bioactivity data import	*JavaScript/Ajax*	tabular data	table/heat map	SAR table
Structure depictions	*CACTVS*	SMILES/SDF	Image file (GIF)	one or many compounds
Structure drawing	*JME Molecular Editor*	mouse clicks	SMILES/SDF	single compound
Database import	*SOAP*	XML/SDF	SMILES/SDF	PubChem
Scriptable access from *R*	*ChemmineR**	SDF, tabular data	online viewing	SAR table
(II) Similarity Toolbox				
Fragment-based similarity	*Atom Pairs**	SDF/SMILES	Similarity coefficients	pairwise comparisons
Maximum common substructure	*MCS**	SDF/SMILES	MCS (SDF), similarity coefficient	pairwise comparisons
(III) Search Toolbox				
Embedding and indexing	*EI Search**	mouse clicks, SDF/SMILES	ranked compound list	database search
Fingerprint search	*PubChem PUG*	mouse clicks, SDF/SMILES	ranked compound list	database search
(IV) Clustering Toolbox				
Binning clustering	*cmp.cluster**	SDF/SMILES, custom table	cluster table	
Hierarchical clustering	*hclust*	SDF/SMILES, custom table	tree, distance matrix	optional heat map
Multidimensional scaling	*cmdscale*	SDF/SMILES, custom table	scatter plot	interactive
(V) Property Toolbox				
Physicochemical descriptors	*JOELib*	SDF/SMILES	property table	38 descriptors

Table 2.1: **List of services provided by *ChemMine Tools*.** The names of software tools, libraries and environments are italicized. Programs developed by the *ChemMine Tools* project are labeled with a star*. Acronyms defined in text.

Here, we present *ChemMine Tools* as an online portal to a variety of cheminformatics, visualization, search and clustering tools for small molecule data. The utilities provided by this service are useful for various analysis and data mining routines of small molecule screening experiments in chemical genomics and related areas. An easy to use web interface makes these tools accessible to experimental scientists without an extensive computational background.

2.3 Methods

Conceptually, the *ChemMine Tools* online service is divided into five application domains (Figure 7.2; Table 2.1): (i) a *Compound Workbench* for data imports and result management; (ii) a *Structure Similarity Toolbox* to quantify the similarities among compounds; (iii) a *Search Toolbox* for retrieving similar compounds from PubChem; (iv) a *Clustering Toolbox* for accessing clustering and data visualization tools; and (v) a *Property Toolbox* for predicting physicochemical properties of compounds. To construct robust data analysis workflows, the back-end of the server employs a modular design architecture with object-oriented methods and container classes assuring compatible input/output flows and parameter settings among the different data processing units. Currently, the server integrates over 30 cheminformatics and data mining tools that were developed by this or related open source projects. The modular organization of the *ChemMine Tools* service has several advantages. For instance, it maximizes the transparency and maintainability of the system, and simplifies the addition of new features and analysis methods upon user request. The web interface of *ChemMine Tools* is written in Python using the object-oriented and highly scalable *Django* web framework. Modern *JavaScript/Ajax* utilities are embedded to generate interactive and customizable high-content web pages. Moreover, the *ChemMine Tools* project is dedicated to an open access and resource sharing policy. All of its online services and downloadable software components are freely available without restrictions. The following subsections give a detailed description of the underlying algorithms and software tools used by the individual *ChemMine Tools* services.

2.4 Discussion of Services

2.4.1 *Compound Workbench*

A central feature of *ChemMine Tools* is its *Compound Workbench.* It provides a flexible online workspace to upload, manage and visualize small molecule data. Compounds can be imported by reading them from local files, copy and paste, *PubChem* queries (see *Search Toolbox*) or by interacting with the service through the *ChemmineR* library [4] within the statistical programming environment *R*. The latter is an extension of the *ChemMine Tools* project to provide a programmable interface to more advanced users. Alternatively, compounds can be drawn online with the *JME Molecular Editor* [12] and then added to the *Compound Workbench.* Currently, the import utility supports the structure data format (SDF) and simplified molecular input line entry system (SMILES). After the import, one can organize and annotate the compounds or view their structure images in single or batch modes. These images are generated in real time from the underlying structure definition data using the structure depiction tool of the *CACTVS* software suite [21] which runs on the server side. To revisit instances of compound sets, users can save their workbench for later use by downloading the compounds to local files. The compound download function also serves as a format conversion tool to interconvert structure representations between SDF and SMILES formats using utilities from the *Open Babel* project [16, 29]. Once the user has populated the *Compound Workbench* with structures, it serves as a central submission system to all downstream analysis services.

2.4.2 *Similarity Toolbox*

In many small molecule screening data analysis routines it is important to compute objective similarity measures among compounds as a means to compare and prioritize structurally related lead compounds. To provide this functionality, *ChemMine Tools* has implemented two algorithms for computing similarity coefficients among compound structures. The first employs atom pairs as structural descriptors [8] and the widely used Tanimoto coefficient as a similarity measure (see below for more details). Alternatively, users can choose other similarity coefficients, such as Tversky or Dice [20]. The second algorithm identifies the maximum common substructure (MCS) shared among compound pairs [5]. Subsequently, the size of both compounds and the size of their shared MCS is used to calculate the available similarity coefficients. The underlying MCS algorithm often provides the most accurate and sensitive similarity measure, especially for compounds with large size differences [33, 18].

2.4.3 *Search Toolbox*

To efficiently mine much of the chemical structure and bioactivity space available in the public domain, the *ChemMine Tools* service provides text and structure similarity search methods that interface with the *PubChem* database [25] via its *SOAP*-based *Power User Gateway* (*PUG*) data exchange feature. During an analysis session, instantaneous search functionality is often important for retrieval of detailed property and annotation information for compounds of interest, or to identify related structures. In *ChemMine Tools*, structural similarity searches can be performed with *PubChem's* fingerprint search engine or via the *EI Search* method. The latter was developed in house as part of this project to provide ultra-fast structure similarity search functionality using an embedding/indexing (EI) algorithm [6]. When the fingerprint method is chosen, the query is sent to *PubChem*, where the structure search is performed and the results are returned to the *compound workbench*. In contrast to this, *EI Search* is specific to the *ChemMine Tools* project and thus, runs locally on its servers. These two tools possess complementary strengths and weaknesses in identifying weak similarities among compounds[6].

2.4.4 *Clustering Toolbox*

Clustering of compounds by structural or property similarity can be a powerful approach to correlating compound features with biological activity. Clustering tools are also widely utilized for diversity analyses to identify structural redundancies and other biases in compound libraries. *ChemMine Tools'* clustering workbench provides an online interface to three clustering algorithms which include hierarchical clustering, multidimensional scaling (MDS) and binning clustering [4]. The following provides a short overview of these tools, while a more detailed outline of the underlying theory and clustering schemes is available in the online tutorial. When clustering by structural similarity, the required similarity measures are computed by first generating the atom pair descriptors (features) for each compound which are then used to calculate a similarity matrix based on the common and unique features observed among all compound pairs using the Tanimoto coefficient. The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. For the subsequent clustering steps, the similarity matrix is converted into a distance matrix by subtracting the similarity values from 1. The hierarchical and MDS clustering methods provided by *ChemMine Tools* are based on the *R* programs *hclust* and *cmdscale*, respectively; the third method utilizes an internally developed C++ implementation. These three programs complement one another with respect to their data outputs and visualization options. Hierarchical clustering organizes compounds by similarity in a tree with branch lengths proportional to the item-to-item (compound-to-compound) similarities, while the MDS output encodes this information in a scatter plot. These two methods do not directly provide assignments of compounds to discrete similarity groups; assignments are generated downstream of the actual clustering process using various post-processing methods, such as tree cutting approaches. The binning clustering output provides these groupings directly for a user-definable similarity cutoff. For instance, if a Tanimoto coefficient of 0.6 is chosen then compounds will be joined into groups that share a similarity of this value or greater using a ‘single linkage’ rule for cluster joining. Final results are presented as interactive visualization pages to simplify the interpretation of the (often complex) clustering results. The hierarchical clustering result page uses the *Google Maps API* to generate zoom- and click-able trees aligned with molecular structure images. Moreover, heat maps of user uploaded data containing compound property, activity or other information can be viewed alongside the tree. A similar system is used to present the MDS results as click-able scatter plots with cursor-over viewing of compound structures. The binning clustering results are presented in a table view containing (among other information) the cluster identifiers and the corresponding compound depictions.

2.4.5 *Property Toolbox*

Predictions of small molecule physicochemical properties are important for assessing their ‘druglikeness’ and ‘leadlikeness’ *in silico* [39, 28]. They are also useful for enriching compound collections with desirable properties. For instance, the famous “Lipinski Rule of Five” [26] is often applied to enrich compound collections with druglike candidates. This rule filters for compounds with ≤ 5 hydrogen bond donors, ≤ 10 hydrogen acceptors, a molecular weight ≤ 500 daltons and an octanol-water partition coefficient log P ≤ 5. Physicochemical property data are essential for predicting bioactive and other properties of small molecules using modern machine learning approaches. These data are fundamental to the development of quantitative structure-activity relationship (QSAR) models [35]. *ChemMine Tools* provides an online interface to the property prediction module of the *JOELib* package [43]. This service can calculate 38 physicochemical property values, including Lipinski descriptors for custom compound sets. The resulting property tables can be downloaded or further processed on *ChemMine Tools* by sending them to the *Clustering Toolbox*. There, they can be used to cluster compounds by similar property profiles, as described above, or the data can be visualized as a heat map next to the hierarchical clustering trees.

2.5 Conclusion and Further Development

ChemMine Tools is an online service for compound analysis in the chemical genomics field. The service is unique in that it integrates a large number of cheminformatic programs with clustering and visualization functionalities. Additional outstanding features of *ChemMine Tools* include: (i) its commitment to publicly developed open source software throughout its infrastructure; (ii) its strong dedication to the development of new cheminformatic tools and their free distribution in the community; and (iii) the integration of its many components into a unified online and downloadable software infrastructure which maximizes their utility for diverse tasks with different levels of complexity and customization needs. An intuitive web interface makes these tools accessible to scientists with limited computational background, while simultaneously providing a programmable interface for advanced users. To the best of our knowledge, there are currently no related online services available that provide a comparable suite of functionalities. Overlaps exist, however they are limited to isolated functionalities. For instance, *ChemDB* and VCCLab [7, 39] can be used for property predictions and structure format interconversions of single compound queries; and *PubChem* supports structure-based clustering for compounds retrieved from its own database.

In the future, many additional utilities will be added to the *ChemMine Tools* service including the addition of MCS-based search functionality within the *Similarity Toolbox* to support more complex graph-based search strategies against custom compound sets imported into the *Compound Workbench.* Existing functionalities for analyzing bioactivity data will also be expanded by adding a *Bioactivity Toolbox* that will contain regression, machine learning and QSAR modeling tools.

2.6 Bibliography

[1] C P Austin, L S Brady, T R Insel, and F S Collins. NIH Molecular Libraries Initiative. *Science*, 306(5699):1138–1139, 2004.

[2] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias K o tter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, 2007.

[3] P Block, C A Sotriffer, I Dramburg, and G Klebe. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res*, 34(Database issue):522–526, 2006.

[4] Y Cao, A Charisi, L C Cheng, T Jiang, and T Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24(15):1733–1734, 2008.

[5] Y Cao, T Jiang, and T Girke. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, 24(13):366–374, 2008.

[6] Y Cao, T Jiang, and T Girke. Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. *Bioinformatics*, 26(7):953–959, 2010.

[7] J H Chen, E Linstead, S J Swamidass, D Wang, and P Baldi. ChemDB update–full-text search and virtual chemical space. *Bioinformatics*, 23(17):2348–2351, 2007.

[8] X Chen and C H Reynolds. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J Chem Inf Comput Sci*, 42(6):1407–1414, 2002.

[9] J Couzin. Molecular medicine. NIH dives into drug discovery. *Science*, 302(5643):218–221, 2003.

[10] K Degtyarenko, P de Matos, M Ennis, J Hastings, M Zbinden, A McNaught, R Alcántara, M Darsow, M Guedj, and M Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 36(Database issue):344–350, 2008.

[11] C M Dobson. Chemical space and biology. *Nature*, 432(7019):824–828, 2004.

[12] P Ertl. Molecular structure input on the web. *J Cheminform*, 2(1):1–1, 2010.

[13] T Girke, L C Cheng, and N Raikhel. ChemMine. A compound mining database for chemical genomics. *Plant Physiol*, 138(2):573–577, 2005.

[14] A Goede, M Dunkel, N Mester, C Frommel, and R Preissner. SuperDrug: a conformational drug database. *Bioinformatics*, 21(9):1751–1753, 2005.

[15] R Guha. Chemical Informatics functionality in R. *J Stat Softw*, 18(8):1–16, 2007.

[16] R Guha, M T Howard, G R Hutchison, P Murray-Rust, H Rzepa, C Steinbeck, J Wegner, and E L Willighagen. The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model*, 46(3):991–998, 2006.

[17] S J Haggarty. The principle of complementarity: chemical versus biological space. *Curr Opin Chem Biol*, 9(3):296–303, 2005.

[18] M Hattori, Y Okuno, S Goto, and M Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc*, 125(39):11853–11865, 2003.

[19] M Hattori, Y Y Okuno, S Goto, and M Kanehisa. Heuristics for chemical compound matching. *Genome Inform*, 14:144–153, 2003.

[20] J D Holliday, N Salim, M Whittle, and P Willett. Analysis and display of the size dependence of chemical similarity coefficients. *J Chem Inf Comput Sci*, 43(3):819–828, 2003.

[21] W D Ihlenfeldt, J H Voigt, B Bienfait, F Oellien, and M C Nicklaus. Enhanced CACTVS browser of the Open NCI Database. *J Chem Inf Comput Sci*, 42(1):46–57, 2002.

[22] J J Irwin and B K Shoichet. ZINC–a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, 45(1):177–182, 2005.

[23] M Kanehisa, S Goto, M Hattori, K F Aoki-Kinoshita, M Itoh, S Kawashima, T Katayama, M Araki, and M Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):354–357, 2006.

[24] M Kuhn, C von Mering, M Campillos, L J Jensen, and P Bork. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res*, 36(Database issue):684–688, 2008.

[25] Q Li, T Cheng, Y Wang, and S H Bryant. PubChem as a public resource for drug discovery. *Drug Discov Today*, 15(23-24):1052–1057, 2010.

[26] C A Lipinski, F Lombardo, B W Dominy, and Feeney P J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, 1997.

[27] T Liu, Y Lin, X Wen, R N Jorissen, and M K Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*, 35(Database issue):198–201, 2007.

[28] A Monge, A Arrault, C Marot, and L Morin-Allory. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol Divers*, 10(3):389–403, 2006.

[29] N M O'Boyle, C Morley, and G R Hutchison. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem Cent J*, 2(1):5–5, 2008.

[30] M M Olah, C G Bologa, and T I Oprea. Strategies for compound selection. *Curr Drug Discov Technol*, 1(3):211–220, 2004.

[31] T I Oprea, A Tropsha, J L Faulon, and M D Rintoul. Systems chemical biology. *Nat Chem Biol*, 3(8):447–450, 2007.

[32] S A Rahman, M Bashton, G L Holliday, R Schrader, and J M Thornton. Small Molecule Subgraph Detector (SMSD) toolkitl. *J Cheminform*, 1(1):12–12, 2009.

[33] J W Raymond and P Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des*, 16(7):521–533, 2002.

[34] K P Seiler, G A George, M P Happ, N E Bodycombe, H A Carrinski, S Norton, S Brudz, J P Sullivan, J Muhlich, M Serrano, P Ferraiolo, N J Tolliday, S L Schreiber, and P A Clemons. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res*, 36(Database issue):351–359, 2008.

[35] O Spjuth, E L Willighagen, R Guha, M Eklund, and J E Wikberg. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J Cheminform*, 2(1):5–5, 2010.

[36] C Steinbeck, C Hoppe, S Kuhn, M Floris, R Guha, and E L Willighagen. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des*, 12(17):2111–2120, 2006.

[37] R L Strausberg and S L Schreiber. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science*, 300(5617):294–295, 2003.

[38] V J Sykora and D E Leahy. Chemical Descriptors Library (CDL): A Generic, Open Source Software Library for Chemical Informatics. *J Chem Inf Model*, 2008.

[39] I V Tetko, J Gasteiger, R Todeschini, A Mauri, D Livingstone, P Ertl, V A Palyulin, E V Radchenko, N S Zefirov, A S Makarenko, V Y Tanchuk, and V V Prokopenko. Virtual computational chemistry laboratory–design and description. *J Comput Aided Mol Des*, 19(6):453–463, 2005.

[40] J H Voigt, B Bienfait, S Wang, and M C Nicklaus. Comparison of the NCI open database with seven large chemical structural databases. *J Chem Inf Comput Sci*, 41(3):702–712, 2001.

[41] T Walker, C M Grulke, D Pozefsky, and A Tropsha. Chembench: a cheminformatics workbench. *Bioinformatics*, 26(23):3000–3001, Dec 2010.

[42] R Wang, X Fang, Y Lu, and S Wang. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem*, 47(12):2977–2980, 2004.

[43] J K Wegner, H Fröhlich, and A Zell. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J Chem Inf Comput Sci*, 44(3):931–939, 2004.

[44] D S Wishart, C Knox, A C Guo, D Cheng, S Shrivastava, D Tzur, B Gautam, and M Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36(Database issue):901–906, 2008.

[45] Q Zhu, M S Lajiness, Y Ding, and D J Wild. WENDI: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J Cheminform*, 2:6–6, 2010.

Chapter 3

bioassayR: Cross-Target Analysis of Small Molecule Bioactivity

3.1 Abstract

Despite a large and rapidly growing body of small molecule bioactivity screens available in the public domain, systematic leverage of the data to assess target druggability and compound selectivity has been confounded by a lack of suitable cross-target analysis software. We have developed bioassayR, a computational tool which enables simultaneous analysis of thousands of bioassay experiments performed over a diverse set of compounds and biological targets. Unique features include support for large-scale cross-target analyses of both public and custom bioassays, generation of high throughput screening fingerprints (HTSFPs), and an optional pre-loaded database which provides access to a substantial portion of publicly available bioactivity data.

bioassayR is implemented as an open-source R/Bioconductor package available from https://bioconductor.org/packages/bioassayR/.

3.2 Introduction

Diverse collections of small molecules have been screened over the past decade against a wide array of distinct protein target families. The resulting high throughout screening (HTS) data are available in community driven databases such as PubChem Bioassay, ChEMBL, ZINC, ChemDB, and many others (list in table S1 of Supporting Information)[41, 12, 21, 6]. As demonstrated by many data mining efforts, these bioactivity resources provide an opportunity for studying the selectivity patterns and molecular mechanisms of small molecule-target interactions on a broad scale[24, 44, 43, 16, 7, 41, 36, 35]. These insights have the potential to lead to the discovery of drug candidates and protein target sites relevant for medical or chemical genomics applications. The data can also be used to identify and exclude drug candidates with largely unselective binding properties (*e.g.* promiscuous binders) which have been found to be of limited use to most application areas [16, 28, 8]. Moreover, the bioactivity data can be used to develop multi-target treatments specific to one or several cross-connected pathways; to identify alternative uses for existing drugs; or to predict potential side and toxic effects [34, 25, 26]. Data from single target screens (*i.e.* a bioassay with a specific target protein) can also be helpful for prioritizing potential target sites in multiplexed or high-content screens, where a specific target protein is usually unknown. Furthermore, large-scale compound bioassay data can be used to create an inventory of molecular functions and proteins that are accessible or resistant to perturbations by small molecules. These “druggability profiles” can be used to guide decision processes in selecting the most efficient target sites for a specific research application in drug discovery and other small molecule driven research disciplines [9].

Most of the small molecule bioactivity data available in the above mentioned public databases were generated by systematic screening efforts of the Molecular Libraries Program (MLP), the Chemical Biology Program of the Broad Institute, and a variety of smaller public efforts [1]. The online interfaces of these databases provide many useful search and download options for focused analysis of a small number of molecules or target proteins[41, 12]. While several projects have developed statistical methods and sample scripts applicable to cross-target analysis, there is currently no general purpose software infrastructure available to perform these tasks in a systematic and fully customizable manner [13, 24, 44, 43, 16, 7, 8, 16].

To address this deficit, we have developed bioassayR, a computational package for the statistical programming language R which enables simultaneous analysis of numerous bioassay experiments performed across diverse compounds and biological targets [32]. bioassayR is distinct from existing tools for analyzing high throughput screening data in several important ways: (i) its focus on the simultaneous tracking and comparative analysis of a large number of assays of distinct experimental design and source; (ii) its flexible data structures optimized for performance with large data and interoperability with existing statistical software; (iii) its integration with numerous R language cheminformatics and bioinformatics tools curated by the Bioconductor and CRAN projects, including ChemmineR, ChemmineOB, rcdk, cellHTS, fmcsR, and eiR [14, 32, 4, 15, 5, 40, 2]. For example, users can analyze their own HTS data (*e.g.* processed with cellHTS) alongside public bioactivity data; or process bioactivity fingerprints (HTSFPs) with functionalities provided by ChemmineR. HTSFPs summarize the activity of compounds across many protein targets. Several studies have demonstrated their effectiveness in predicting and categorizing bioactivity in a manner complementing rather than overlapping with structure based predictions [18, 30, 42, 33, 22, 23, 11]. In addition, they can be used as trainings data sets for predicting active ligand-target pairs with supervised machine learning algorithms [27, 37]. The HTSFP tools implemented in bioassayR will generate fingerprints for any custom set of compounds and targets, optionally merge assays with similar or identical targets, and compare activity profiles by either continuous z-scores or binary active/inactive values. Z-score based HTSFPs exhibit greater predictive power in hit expansion experiments, while binary HTSFPs require less computational overhead, enabling all-against-all bioactivity profile comparison for hundreds of thousands of compounds [33].

3.3 Methods and Implementation

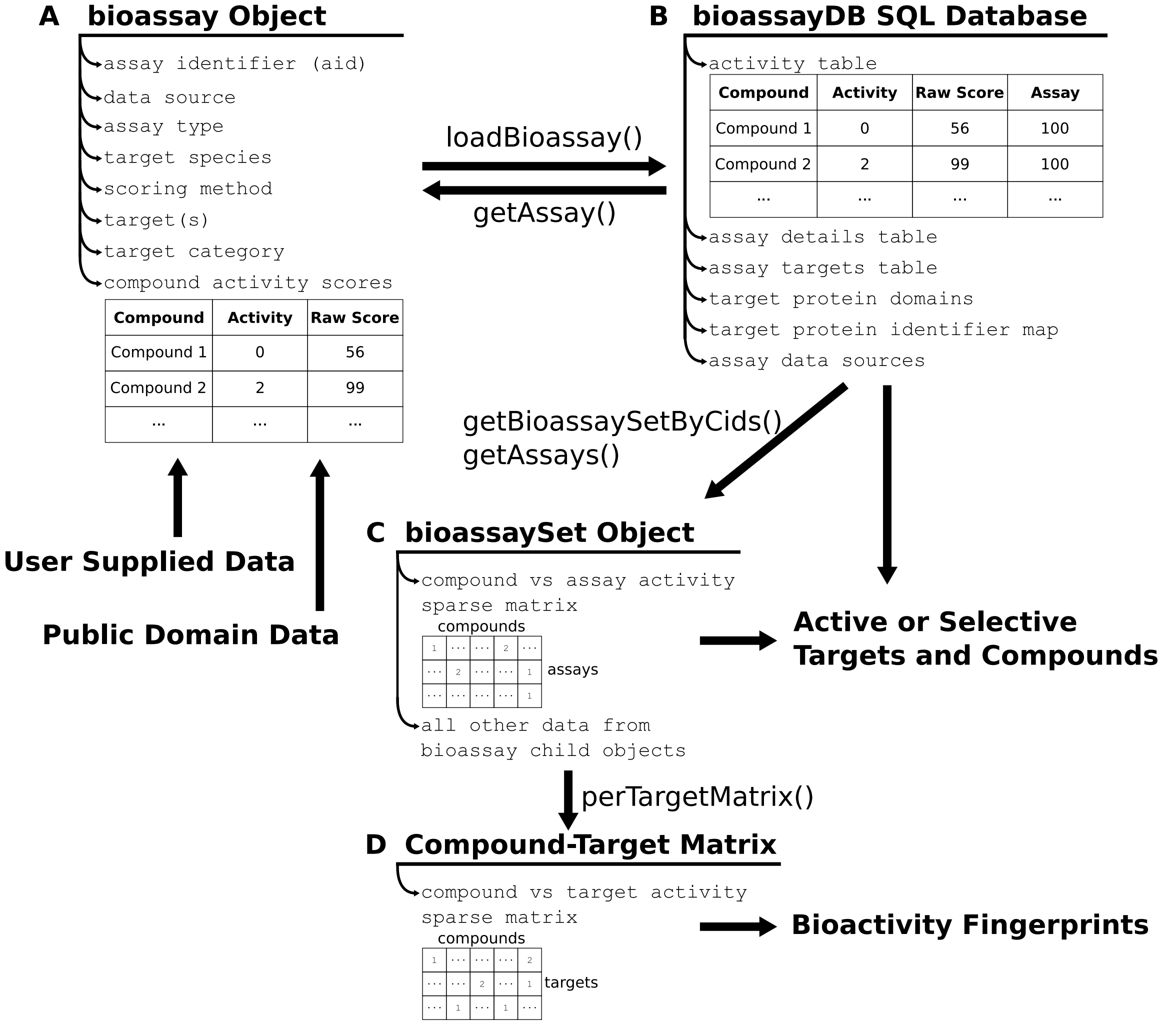
3.3.1 Software design and workflow overview

bioassayR's data model is designed around four interconnected data objects (R language S4 classes), each with an internal structure optimized for different bioactivity analysis routines. They are introduced below in more detail and Figure 3.1 provides an illustration. In short, the bioassay data is organized in an SQL database called *bioassayDB*; data from single and many assays are imported into *bioassay* and *bioassaySet* objects, respectively; and the compound-target matrix summarizes the compound vs. target activities from many assays. The *bioassayDB* serves as a large data repository which can efficiently organize and query millions of assays simultaneously, while the other objects facilitate analysis of a subset of these data selected to answer a specific biological question. Table S2 in Supporting Information lists selected cross-target analysis functions which query the data within these objects. Users can optionally use a pre-built *bioassayDB* database which contains publicly available bioactivity data against a wide range of protein targets.

3.3.2 bioassay object: importing data

The *bioassay* object (Figure 3.1 section A) stores data from a single bioassay experiment, and acts as a gateway for importing new assay data, as well as for editing and investigating data from one assay at a time. This object stores the assay identifier (aid), data source, assay type, target species, scoring method, target identifiers, target categories, and activity scores.

bioassayR provides users with the option of performing analyses either on their own bioactivity data, on a prebuilt database of public domain bioactivity data, or both simultaneously. Four options exist for importing data as a bioassay object: (i) data in the standard PubChem CSV and XML formats can be parsed with a built in function;



Compound	Activity	Raw Score
Compound 1	0	56
Compound 2	2	99
...	...	...

Compound	Activity	Raw Score	Assay
Compound 1	0	56	100
Compound 2	2	99	100
...	...	...	...

1	⋯	⋯	2	⋯
⋯	2	⋯	⋯	1
⋯	⋯	⋯	⋯	1

all other data from bioassay child objects
Active or Selective Targets and Compounds
perTargetMatrix()
D Compound-Target Matrix
compound vs target activity sparse matrix
compounds
targets

1	⋯	⋯	⋯	2
⋯	⋯	2	⋯	1
⋯	1	⋯	1	⋯

Figure 3.1: **Design overview and workflow.** bioassayR stores bioactivity data in four interconnected objects. **A** Data from a single bioassay experiment is imported into a bioassay object. **B** Any number of bioassay objects can be loaded into the bioassayDB SQL database which is optimized for time efficient searching. **C** Filter and query methods are available to identify compounds or assays of interest. These query results can be imported into a bioassaySet object which stores activity data as a sparse matrix where columns represent compounds and rows assays (targets). This organization facilitates many typical cross-target analysis routines, *e.g.* target selectivity analyses. **D** To reduce both redundancy and sparseness in the data, assays involving the same or similar targets can be collapsed into a single row using the *perTargetMatrix* function.

(ii) data already represented as an R *data.frame* or tabular file with activity values can be directly converted into a bioassay object; (iii) raw screening data from a microtiter plate reader can be analyzed using the cellHTS2 R package, and converted into a bioassay object; (iv) extracting a single assay from an already existing bioassayR database, such as the prebuilt PubChem BioAssay database described below[2]. All four options are demonstrated with examples in the package documentation. Once represented as a bioassay object, these data can be viewed, edited, or loaded into a *bioassayDB* database for analysis alongside other assays.

3.3.3 bioassayDB object: multiple assay SQL database

The *bioassayDB* object (Figure 3.1 section B) stores a connection to a SQL database optimized for efficient aggregate search-based analysis across multiple assays. Users can load, edit, or delete individual *bioassay* objects, and then query these data. Many analysis and query functions are provided to investigate the data within a *bioassayDB* object (see Table S2 in Supporting Information). The database is contained within a single file that can be easily shared among users. Internally, the database stores data from a large number of individual bioassay objects, in addition to target protein domain data, and target identifier mappings. Multiple types of identifier mapping and annotation data can be stored, for example to translate target identifiers into those used by common databases such as UniProt, or to annotate proteins by storing categorization data such as a sequence-similarity clustering bin for each protein [39].

3.3.4 bioassaySet object: storing multiple assays in a matrix

Query results from *bioassayDB* can be stored as a *bioassaySet* (Figure 3.1 section C). This matrix-like object along with its accessor methods abstracts complicated analysis tasks across large numbers of compounds and bioassays. By representing bioactivity data as a compound vs. assay matrix, the full range of matrix operations in R can be leveraged to analyze these data efficiently. For example, rows can be compared to compute the similarity between the activity profiles of two molecules. Sparse matrix compression is utilized to avoid unnecessary usage of system memory by untested compound-target combinations. In a typical workflow, a user will first query the database to find a list of compounds or assays of interest, and then extract these into a *bioassaySet* for further analysis.

To address questions of compound vs. target bioactivity, bioassayR can transform a *bioassaySet* into a compound-target matrix by merging assays which share common or similar target proteins, such as close orthologs from different species. Replicates and similar-target assays can be summarized into single values by either specifying a custom summary statistic, or choosing among several provided. The compound vs. target matrix can be generated from either discrete “active” or ”inactive” activity categories, or from continuous activity scores to serve as either binary or continuous numeric HTSFPs, respectively. The *scaleBioassaySet* function will scale and center continuous scores to create a z-score fingerprint. Optionally, omitting inactive values from the discrete activity categories will produce a matrix suitable for analysis with binary matrix algorithms. This data structure can serve as a bipartite graph (or bigraph) connecting compounds and targets, allowing users to analyze these data with the numerous graph and network analysis algorithms available for the R programming language.

3.3.5 Pre-built PubChem BioAssay database

To enable efficient analyses across large numbers of compounds and protein targets, we provide downloadable instances of the *bioassayDB* database pre-loaded with public bioactivity data. This frequently updated database file includes all screens from PubChem BioAssay involving known target proteins. PubChem BioAssay data has been chosen since it includes assays from many sources such as ChEMBL, and therefore represents a substantial portion of all publicly available bioactivity data. At the time of this writing the data contains activity results from roughly 1.2 million structurally distinct compounds tested against protein 6,339 targets. As many compound-target combinations have not been tested, these data are sparse with roughly half (572,947) of the compounds having screening results for at least 10 distinct protein targets. Among these “highly screened" compounds, 895 are currently FDA approved drugs. PubChem BioAssay provides bioactivity data both as continuous numeric scores, and active/inactive categories.

To extend the utility of these data, we provide and include within the pre-built database additional annotation details for each protein target. The database includes both NCBI Protein GI numbers and UniProt identifiers for all protein targets, Pfam domains identified with the HMMER software, and amino acid sequence similarity-based clustering performed with kClust[39, 31, 10, 17]. The UniProt identifiers allow users to obtain further annotation details including Gene Ontology (GO) terms programmatically by connecting to external annotation databases [38]. The Pfam domain mappings provide groupings for local similarities and across wider evolutionary distances, while the sequence similarity cluster are more suitable for identifying groups of sequences sharing a defined degree of sequence similarity.

The included annotation data expand the usefulness of bioassayR for several applications. For instance, the annotations can be used for merging similar assays into a compound-target matrix as described in the above “bioassaySet object” section. When searching for compounds active against a desired protein, users can expand the search to include compounds found active against protein targets that share sequence similarity, domains, or GO terms with the query. This method can identify compounds that are likely active against a target of interest, even if little or no screening data exists for that specific target. In drug discovery experiments where a specific protein target has not yet been identified, these data can help identify protein targets worth investigating based on presence of a specific protein domain, molecular function, or orthologue that has been previously found to be involved in the desired therapeutic effect.

3.3.6 Identifying compounds with selective or promiscuous bioactivity

Bioactive small molecules can be classified according to the quantity of distinct molecular targets they are active against. Target selective compounds bind to a small number of target proteins, whereas “promiscuous binders", indiscriminately bind to a large number of targets. Patterns of target selectivity in widely used drugs can also be used as a template for identifying drug candidates with similar selectivity profiles.

Several bioassayR functions facilitate identification of target selective compounds and the reverse, compound selective targets, across a large set of bioassay experimental results. The *targetSelectivity* function will return the target selectivity for a query compound. To find compounds active against a target or a set of targets in a pathway of interest, the function *activeAgainst* will return all active compounds, while *selectiveAgainst* will return only compounds most selective against the specified target, along with a corresponding selectivity score for each. To consider only compounds that have been tested in numerous assays, the *screenedAtLeast* function will identify compounds that have participated in a specified minimum quantity of screens. To find all targets of a query compound, the functions *activeTargets* and *inactiveTargets* will return the list of active and inactive targets, respectively. The *crossReactivityProbability* function uses a beta-binomial statistical model to estimate the probability that a given compound is a promiscuous binder [8].

3.3.7 Clustering small molecules by bioactivity profile

With bioassayR, large-scale screening data can be used to cluster small molecules based on the similarity of their bioactivity profiles across many target proteins. To cluster small molecules by bioactivity, it is necessary to choose an appropriate similarity measure, such as correlation coefficients which are appropriate for continuous activity data, and the Jaccard or Tanimoto coefficient for categorical or binary data [19]. Next, the chosen similarity measure is used to compute a distance matrix (d) for all possible pair-wise comparisons of bioactivity profiles, by subtracting the similarity values (s) from one: $d = 1 - s$. The distance matrix can then be used as direct input to a variety of clustering algorithms, including hierarchical clustering, k-means or multiple dimensional scaling (MDS).

The bioassayR clustering workflow starts by generating a compound-target bioactivity matrix, as described above, with either continuous or discrete category activity scores. For continuous scores, several similarity functions available in R, such as the base function *cor* can be used to create a distance matrix based on Pearson correlation coefficients. The associated ChemmineR package will create a distance matrix for binary bioactivity fingerprints generated by bioassayR. Comparisons among binary ChemmineR fingerprints have less CPU and memory overhead than continuous z-score based comparisons, and therefore are suitable for all-against-all comparisons of larger compound sets.

By default, the bioassayR HTSFPs features resolve missing (untested) activity values by assuming inactivity, where a ‘0’ is used for binary fingerprints, and a z-score of ‘0’ is used for continuous fingerprints. When computing the similarity between two compound bioactivity profiles, this can lead to false negatives (lower than the true similarity value) if the compounds share few common screened targets [33]. A more accurate estimate of similarity can be obtained by using machine learning methods which impute the missing values, however this introduces false positives which are often less desirable than false negatives in drug discovery efforts [33]. The bioassayR function *screenedAtLeast* can limit false negatives without introducing false positives by including only highly screened compounds in the analysis. Alternatively, the compound vs. target matrix can be subset with a biclustering algorithm to limit similarity comparison to a densely screened subset of a larger sparse compound vs. target matrix. Lastly, bioassayR also provides a similarity function (*trinarySimilarity*) which avoids assuming inactivity for missing compound-target activity values by operating on a trinary bioactivity matrix which uses a ‘0’ for untested or missing values, a ‘1’ for inactive values, and a ‘2’ for active values. This function computes similarity based only on the mutually screened targets between two compounds, and returns an “NA” if insufficient shared assays exist to make a meaningful comparison. The strategy of performing the comparison only on mutually screened targets, with a minimum threshold for informative data was inspired by the continuous score “Assay Performance Profile Similarity” metric published by Dančík, V *et al.* [8]

3.4 Results and Discussion

In Supporting Information, we highlight three example use cases demonstrating the utility of bioassayR. First, we investigate the diversity of public screening data provided by PubChem BioAssay, and show that these data contain compounds active against a large number of novel protein targets that are not currently accessible with FDA approved drugs. Second, we use bioassayR to cluster FDA approved drugs by bioactivity profiles as well as molecular structure to demonstrate that many drugs exhibit distinct bioactivity patterns that cannot be inferred from structure alone. Third, we demonstrate how bioassayR can be used to enrich a screening library with active compounds and how to guide the time consuming target site identification processes in high-content screening. The vignette (user manual) of the package contains additional examples including loading custom screening data, identifying target selective compounds, and performing custom database queries.

It is important to point out that HTS data are noisy and error prone due to several causes including experimental noise, and incorrect annotation. While public bioactivity databases have implemented strategies to identify and reduce errors, we caution bioassayR users to expect some level of error and mis-annotation depending on the source and type of data used [29]. The impact of these errors on analysis results can be minimized by incorporating replicates and confirmatory screening results from different sources using the bioassayR functions described above. If appropriate, error can also be reduced by limiting analysis to the subset of public bioactivity data that has been manually curated and carefully annotated with a machine readable, non-ambiguous structured vocabulary from sources such as the BioAssay Research Database (BARD) [20, 35].

The bioassayR package is a flexible computational environment for simultaneous analysis of large numbers of high-throughput small molecule bioactivity screens. By organizing large bioactivity data for rapid access and manipulation within the R programming language, bioassayR leverages the substantial breadth of these data as a reference to identify regions of the genome and proteome accessible to small molecule probes, elucidate mechanisms of action for bioactive molecules, and identify off-target effects which currently lead to a high attrition rate in drug discovery efforts [3]. bioassayR provides features to inform the design and analysis of bioactivity and drug discovery experiments; for example to build compound libraries enriched for a desired bioactivity, reducing the search space for effective drugs, druggable protein targets, and chemical genetic probes. bioassayR has functions to identify compounds that have demonstrated activity against targets and pathways of interest, or other targets with sequence or annotation similarity to targets of interest. To build drug discovery libraries with reduced chances of off-target effects, bioassayR will rank compounds for selectivity against a desired target and exclude compounds which show activity against a large number of other targets. To identify compounds or combinations of compounds likely to exhibit a desired polypharmacology (activity against multiple targets), bioassayR will identify all active compounds among a set of query targets. To assess the potential druggability of protein targets, bioassayR will report the quantity and target selectivity of known active drugs and other compounds. To identify compounds with activity similar to existing drugs or other compounds with a known utility, the HTSFPs features enables clustering by cross-target activity profiles. Custom screening data can also be analyzed side-by-side with public data to study the selectivity profiles among newly identified actives across numerous targets, or to assess the level of agreement with any public data that the custom assay replicates.

In addition to providing numerous analysis functions, bioassayR also serves as a bridge to facilitate analysis of large screening data with other machine learning, statistical inference, network analysis, and bioinformatics tools. Many of these tools support the output formats produced by bioassayR with little or no changes. In conclusion, bioassayR lowers the barrier to address questions related to the target selectivity of small molecules with large-scale bioactivity data.

3.4.1 Abbreviations

HTSFPs, High Throughput Screening Fingerprints; HTS, High Throughput Screening

3.5 Bibliography

[1] Christopher P Austin, Linda S Brady, Thomas R Insel, and Francis S Collins. NIH Molecular Libraries Initiative. *Science*, 306(5699):1138–1139, November 2004.

[2] Michael Boutros, Lígia P Brás, and Wolfgang Huber. Analysis of cell-based RNAi screens. *Genome Biol.*, 7(7):R66, 2006.

[3] Joanne Bowes, Andrew J Brown, Jacques Hamon, Wolfgang Jarolimek, Arun Sridhar, Gareth Waldron, and Steven Whitebread. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discovery*, 11(12):909–922, November 2012.

[4] Y Cao, A Charisi, L C Cheng, T Jiang, and T Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24(15):1733–1734, July 2008.

[5] Y Cao, T Jiang, and T Girke. Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. *Bioinformatics*, 26(7):953–959, March 2010.

[6] J Chen, S J Swamidass, Y Dou, J Bruand, and P Baldi. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, 21(22):4133–4139, November 2005.

[7] T Cheng, Y Wang, and S H Bryant. Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules. *Bioinformatics*, 26(22):2881–2888, November 2010.

[8] V Dan ik, H Carrel, N E Bodycombe, K P Seiler, D Fomina-Yadlin, S T Kubicek, K Hartwell, A F Shamji, B K Wagner, and P A Clemons. Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screening*, 19(5):771–781, May 2014.

[9] Sivaraman Dandapani and Lisa A Marcaurelle. Grand Challenge Commentary: Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.*, 6(12):861–863, December 2010.

[10] Sean R Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, 23(1):205–211, October 2009.

[11] A F Fliri, W T Loging, P F Thadeio, and R A Volkmann. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.*, 102(2):261–266, January 2005.

[12] A Gaulton, L J Bellis, A P Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and J P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(D1):D1100–D1107, December 2011.

[13] Peter Gedeck, Bernhard Rohde, and Christian Bartels. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.*, 46(5):1924–1936, September 2006.

[14] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.

[15] Rajarshi Guha. Chemical Informatics Functionality in R. *J. Stat. Softw.*, 18(1):1–16, January 2007.

[16] L Han, Y Wang, and S H Bryant. A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics*, 25(17):2251–2255, August 2009.

[17] Maria Hauser, Christian E Mayer, and Johannes Söding. kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinf.*, 14(1):248, 2013.

[18] Kazi Yasin Helal, Mateusz Maciejewski, Elisabet Gregori-Puigjané, Meir Glick, and Anne Mai Wassermann. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J. Chem. Inf. Model.*, 56(2):390–398, February 2016.

[19] John D Holliday, Naomie Salim, Martin Whittle, and Peter Willett. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.*, 43(3):819–828, May 2003.

[20] E A Howe, A de Souza, D L Lahr, S Chatwin, P Montgomery, B R Alexander, D-T Nguyen, Y Cruz, D A Stonich, G Walzer, J T Rose, S C Picard, Z Liu, J N Rose, X Xiang, J Asiedu, D Durkin, J Levine, J J Yang, S C Schurer, J C Braisted, N Southall, M R Southern, T D Y Chung, S Brudz, C Tanega, S L Schreiber, J A Bittker, R Guha, and P A Clemons. BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.*, 43(D1):D1163–D1170, January 2015.

[21] John J Irwin and Brian K Shoichet. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45(1):177–182, January 2005.

[22] Lawrence M Kauvar, Deborah L Higgins, Hugo O Villar, J Richard Sportsman, Åsa Engqvist-Goldstein, Robert Bukar, Karin E Bauer, Hara Dilley, and David M Rocke. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol. (Oxford, U. K.)*, 2(2):107–118, February 1995.

[23] Lawrence M Kauvar, Hugo O Villar, J Richard Sportsman, Deborah L Higgins, and Donald E Schmidt Jr. Protein affinity map of chemical space. *J. Chromatogr. B Biomed. Sci. Appl.*, 715(1):93–102, September 1998.

[24] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 25(2):197–206, February 2007.

[25] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, 6:343, 2010.

[26] Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Côté, Brian K Shoichet, and Laszlo Urban. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367, June 2012.

[27] Alessandro Lusci, Michael Browning, David Fooshee, Joshua Swamidass, and Pierre Baldi. Accurate and efficient target prediction using a potency-sensitive influence-relevance voter. *J. Cheminf.*, 7:63, 2015.

[28] Susan L McGovern, Brian T Helfand, Brian Feng, and Brian K Shoichet. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.*, 46(20):4265–4272, September 2003.

[29] George Papadatos, Anna Gaulton, Anne Hersey, and John P Overington. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput.-Aided Mol. Des.*, 29(9):885–896, July 2015.

[30] Paula M Petrone, Benjamin Simms, Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhan Deng, John W Davies, Jeremy L Jenkins, and Meir Glick. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.*, 7(8):1399–1409, August 2012.

[31] M Punta, P C Coggill, R Y Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, K Forslund, G Ceric, J Clements, A Heger, L Holm, E L L Sonnhammer, S R

Eddy, A Bateman, and R D Finn. The Pfam protein families database. *Nucleic Acids Res.*, 40(D1):D290–D301, December 2011.

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[33] Sereina Riniker, Yuan Wang, Jeremy L Jenkins, and Gregory A Landrum. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.*, 54(7):1880–1891, July 2014.

[34] U Schmidt, S Struck, B Gruening, J Hossbach, I S Jaeger, R Parol, U Lindequist, E Teuscher, and R Preissner. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.*, 37(Database):D295–D299, January 2009.

[35] S C Schurer, U Vempati, R Smith, M Southern, and V Lemmon. BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. *J. Biomol. Screening*, 16(4):415–426, April 2011.

[36] C Senger, B A Gruning, A Erxleben, K Doring, H Patel, S Flemming, I Merfort, and S Gunther. Mining and evaluation of molecular relationships in literature. *Bioinformatics*, 28(5):709–714, February 2012.

[37] Othman Soufan, Wail Ba-alawi, Moataz Afeef, Magbubah Essack, Valentin Rodionov, Panos Kalnis, and Vladimir B Bajic. Mining Chemical Activity Status from High-Throughput Screening Assays. *PLoS ONE*, 10(12):e0144426, December 2015.

[38] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 43(D1):D1049–D1056, January 2015.

[39] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 40(D1):D71–D75, December 2011.

[40] Y Wang, T W H Backman, K Horan, and T Girke. fmcsR: mismatch tolerant maximum common substructure searching in R. *Bioinformatics*, 29(21):2792–2794, October 2013.

[41] Y Wang, J Xiao, T O Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, B A Shoemaker, E Bolton, A Gindulyte, and S H Bryant. PubChem's BioAssay Database. *Nucleic Acids Res.*, 40(D1):D400–D412, December 2011.

[42] Anne Mai Wassermann, Eugen Lounkine, Laszlo Urban, Steven Whitebread, Shanni Chen, Kevin Hughes, Hongqiu Guo, Elena Kutlina, Alexander Fekete, Martin Klumpp, and Meir Glick. A Screening Pattern Recognition Method Finds

New and Divergent Targets for Drugs and Natural Products. *ACS Chem. Biol.*, 9(7):1622–1631, July 2014.

[43] Anne Mai Wassermann, Lisa Peltason, and Jürgen Bajorath. Computational Analysis of Multi-target Structure-Activity Relationships to Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem*, 5(6):847–858, April 2010.

[44] S Frank Yan, Frederick J King, Yun He, Jeremy S Caldwell, and Yingyao Zhou. Learning from the Data: Mining of Large High-Throughput Screening Databases. *J. Chem. Inf. Model.*, 46(6):2381–2395, November 2006.

Chapter 4

bioassayR: Supporting Information

Name (alphabetical)	Url
BindingDB[11]	`http://www.bindingdb.org`
BioAssay Research Database (BARD)[8]	`https://bard.nih.gov`
CellMiner[15]	`http://discover.nci.nih.gov/cellminer/`
ChemBank[14]	`http://chembank.broadinstitute.org`
ChEMBL[5]	`https://www.ebi.ac.uk/chembldb/`
ChemDB[3]	`http://cdb.ics.uci.edu`
ChemMine[6]	`http://chemminedb.ucr.edu`
DrugBank[19]	`http://www.drugbank.ca`
PubChem[17, 18]	`http://pubchem.ncbi.nlm.nih.gov`
Structural Biology Knowledgebase[4]	`http://sbkb.org`
ZINC[9, 10]	`http://zinc.docking.org`

Table 4.1: **Small molecule databases.** Here we provide links to several small molecule structure and/or bioactivity databases that can potentially be used with bioassayR. This list is not comprehensive.

Function Name	Description
activeAgainst	Returns the active compounds for a query protein target
activeTargets	Returns the active protein targets for a query compound
inactiveTargets	Returns the inactive protein targets for a query compound
bioactivityFingerprint	Creates ChemmineR compatible binary bioactivity fingerprints (HTSFPs) for a list of compounds and targets [2]
crossReactivityProbability	Uses a Bayesian model to compute the probability that a compound is a promiscuous binder
crossReactivityPrior	Computes a prior distribution of binding promiscuity for use with crossReactivityProbability
getBioassaySetByCids	Creates a *bioassaySet* compound vs. assay matrix for a list of query compounds
getAssays	Creates a *bioassaySet* compound vs. assay matrix for a list of query assays
perTargetMatrix	Creates a compound vs. target activity matrix merging replicates as specified
queryBioassayDB	Runs a custom SQL query on a *bioassayDB* bioactivity database
screenedAtLeast	Returns a list of compounds screened against at least a specified number of distinct targets
selectiveAgainst	Returns compounds most selective for a query target
targetSelectivity	Returns the target selectivity for a set of query compounds
translateTargetId	Returns UniProt identifiers or other annotation details for a query target protein
trinarySimilarity	Compute similarity between bioactivity profiles by considering only commonly screened targets
scaleBioassaySet	Centers and/or scales continuous numeric activity scores, creating per-assay z-scores by default

Table 4.2: **Selected cross-target analysis functions.** Instructions and examples are provided in the online package documentation available at http://bioconductor.org/packages/bioassayR/.

4.1 Use case examples

The following presents three use cases illustrating bioassayR's utility for performing cross-target analysis tasks. The vignette (user manual) of the package, available on the Bioconductor site, contains additional examples including loading custom screening data, identifying target selective compounds, and performing custom database queries. For some assays cross-target analyses have to be limited to activity categories because continuous scores are not provided, or were obtained in a unique manner not directly comparable to other assays. For this reason, we performed the examples shown here with activity category data, however, for assays where suitable data is available, analysis can be performed on continuous data.

4.1.1 Drug-target space analysis

To assess the diversity of the protein target space represented by PubChem BioAssay, we enumerated its small molecules and protein targets with bioassayR. Small molecules were grouped into FDA approved drugs and all other compounds, and protein targets were organized by GO slim terms of the Molecular Function Ontology (Figure 4.1) [16]. The PubChem BioAssay data was used to produce a compound-target matrix as described in the "bioassaySet object" section of the manuscript. To highlight the relative distribution of potentially novel compounds in the data, the GO slim terms are ordered in Figure 4.1 according to decreasing numbers of active FDA approved drugs.

As expected, a larger fraction of FDA approved drugs exhibit activity against proteins annotated with GO slim terms than is the case for non-drug compounds. Importantly, a substantial number of active non-drug compounds has been identified within each GO category. Some categories (such as GTPase) have very few FDA approved active drugs, but a large quantity of active non-drug compounds. The FDA approved drugs show activity against 1789 protein targets, whereas the non-drug compounds show activity against an additional 3020 protein targets. Of these

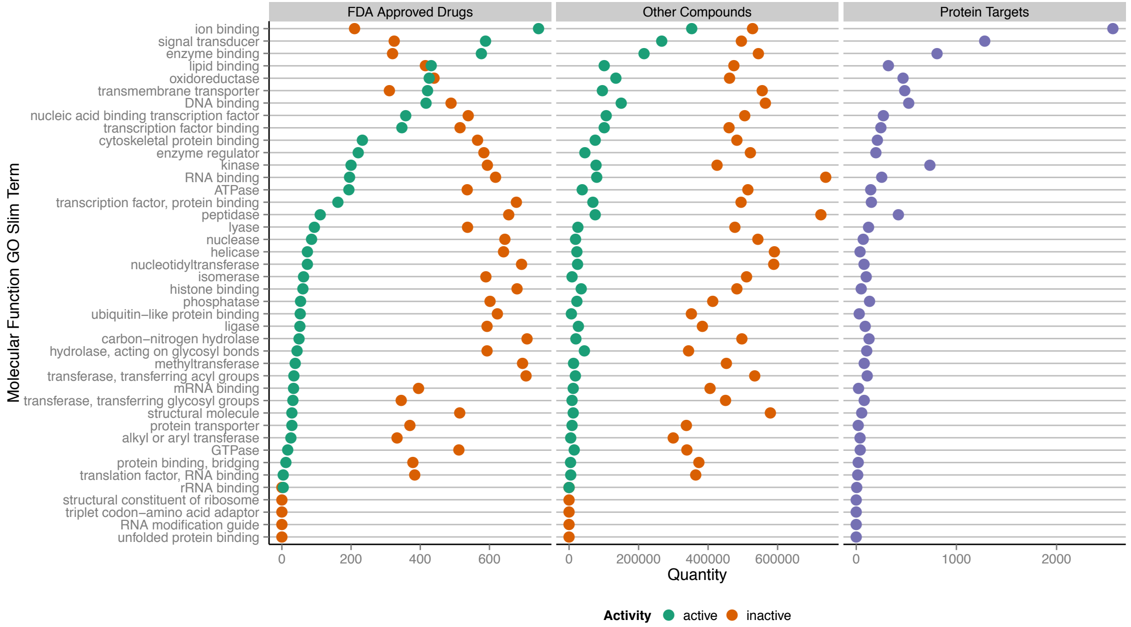


Figure 4.1: **Frequency of active PubChem BioAssay compounds across GO slim categories.** The target proteins represented in PubChem BioAssay have been classified by GO slim terms of the Molecular Function Ontology (vertical axis). The number of compounds with activity or inactivity against at least one target protein within each protein class is plotted on the horizontal axis. The plot on the left gives the numbers for FDA approved drugs, the one in the center for all other compounds, and the one on the right gives the numbers of the corresponding protein targets included in the assays. To distinguish between small and zero quantities of active compounds, the active point is omitted for GO slim terms which have no active compounds.

3020 additional targets, 901 are substantially distinct both from one another and any of the FDA approved drug targets at the amino-acid sequence level based on the kClust clustering described in the "Pre-built PubChem BioAssay database" section of the manuscript. This is a marked expansion of the protein target space currently covered by FDA approved drugs, illustrating the utility of these data for identifying drug candidates with unique properties potentially relevant for the development of medical treatments.

4.1.2 Clustering drugs by structure and bioactivity profile

Using the bioactivity clustering features described in the manuscript, we used bioassayR to cluster all FDA approved drugs screened against at least 10 distinct targets in the PubChem BioAssay data. For the resulting 895 compounds meeting this requirement, we generated a binary HTSFP with the *bioactivityFingerprint* function, with a length of 2019 bits, each representing a screened protein target. While bioassayR supports continuous z-score HTSFPs, binary HTSFPs were chosen for this example to incorporate data from assays provided with activity categories but without continuous numeric scores. ChemmineR was then used to produce a distance matrix based on the Tanimoto similarity coefficient. The distances were used to project each compound into two dimensional plane with MDS (Figure 4.2) where the points are spaced proportionally to the similarity among the bioactivity profiles of each compound [7]. A small random position jitter was applied to minimize overlap of closely spaced points. To assess whether the compounds within bioactivity clusters share structural similarities, the same compounds have also been clustered by structural similarity with functionalities provided by ChemmineR. This structural clustering employed atom pair fingerprints for structure comparisons, the Tanimoto coefficient as similarity metric and single linkage hierarchical clustering with tree cutting at 0.6 to partition the compounds into discrete similarity groups.

The results in Figure 4.2 show several discrete groupings of drugs by bioactivity, while others appear more evenly dispersed across the plane. To compare the results with the outcome of the structure-based clustering, a subset of the structural clusters have been highlighted with distinct colors. Dots with identical colors in close proximity identify groups of compounds sharing both similar structure and bioactivity patterns. However, exceptions are common, where compounds of similar structure are not located in close proximity. This separation can be due to two main reasons: (i) true biological differences in their target binding profiles and/or (ii) a deflation of the similarity coefficients of compound pairs caused by sparseness in their bioactivity profiles. The latter false negatives are a common problem in bioactivity data that has been discussed in the manuscript as well as in previous studies [13]. The results in Figure 4.2 illustrate how analyzing bioactivity data with bioassayR provides additional information that cannot reliably be inferred from structural similarity alone. This includes the systematic identification of both structurally distinct compounds with similar bioactivity profiles, as well as compounds with similar structure but distinct bioactivity.

4.1.3 Guiding high-content screens with reference data from single target bioassays

Single protein target bioassay data can serve as a reference for guiding and interpreting various aspects of high-content screens. For instance, they can be used to assemble small molecule screening libraries enriched in bioactives targeting a biological process of interest. This can be important for low-throughput assays where testing a large number of compounds is not an option. In addition, single target bioassay data can guide the often very time consuming process of identifying target sites in high-content screens. The following illustrates with an example how bioassayR can facilitate these tasks.

The high-content screen chosen for this example is PubChem BioAssay #465 entitled "*Primary HTS assay for chemical inhibitors antigen receptor-induced NF-kappaB activation*". It uses a luciferase reporter to measure NF-kappaB activation in HEK-293 cells. This screen includes bioactivity data for 61,605 compounds. It was chosen as a representative among 345 published high-content bioassays containing at least 10,000 tested compounds.

To identify single target PubChem Bioassay data relevant for this screen, we searched the Reactome database for all *Homo sapiens* entries annotated with either "NF-kB" or "NFkB" [12]. This query returned 1,120 proteins. Of these, we used bioassayR to identify 605 proteins with single target screening data in the PubChem BioAssay database. A binary bioactivity matrix was generated from this data using the bioassayR functions *getBioassaySetByCids* and *perTargetMatrix*. One dimension of this matrix represented the 61,605 compounds from bioassay #465 and the other dimension the single target bioactivity data from the 605 proteins associated with NF-kappaB. We then summed the matrix to obtain the number of active targets for each compound.

The high-content assay reported 128 actives out of a total of 61,605 compounds (0.21% activity). Among the single-target assays, 1,749 out of the same 61,605 compounds were found to be active against one to five NF-kappaB related targets. Table

High-Content PubChem	[1]Total	[2]Active in N Single Target Assays					
BioAssay #465		≥1	1	2	3	4	5
N Compounds	61,605	1,749	1,276	447	19	5	2
N Actives	128	73	47	17	5	3	1
Percent Actives	0.21	4.12	3.68	3.8	26.32	60.0	50.0

Table 4.3: **Results of a high-content screen compared to single target bioassays of the same pathway.** [1]The numbers of total and active compounds are given for the high-content PubChem BioAssay #465. This assay screened for inhibitors of antigen receptor-induced NF-kappaB activation. [2]The same numbers are given for the subset of compounds that have also been found active in single target assays where the corresponding target protein is associated with NF-kappaB processes. Currently, 605 single target assays meet this requirement. The results are subdivided into the number of compounds active in variable numbers of single target assays including ≥1 and 1-5.

4.3 summarizes the distribution. Among these 1,749 single-target active compounds, 73 were also active in the high-content assay (4.17% activity).

This example demonstrates that including the active compounds from relevant single target assays can substantially enrich high-content screening libraries for active compounds, and thus reduce the total number of compounds that need be tested in primary screening efforts. Additionally, single target bioactivity data provides valuable information as it identified candidate target site(s) for 73 of the 128 actives (57%) present in bioassay #465. This information can reduce time- and labor-intensive target identification processes that are often an integral part of many high-content screening efforts conducted in chemical genomics.

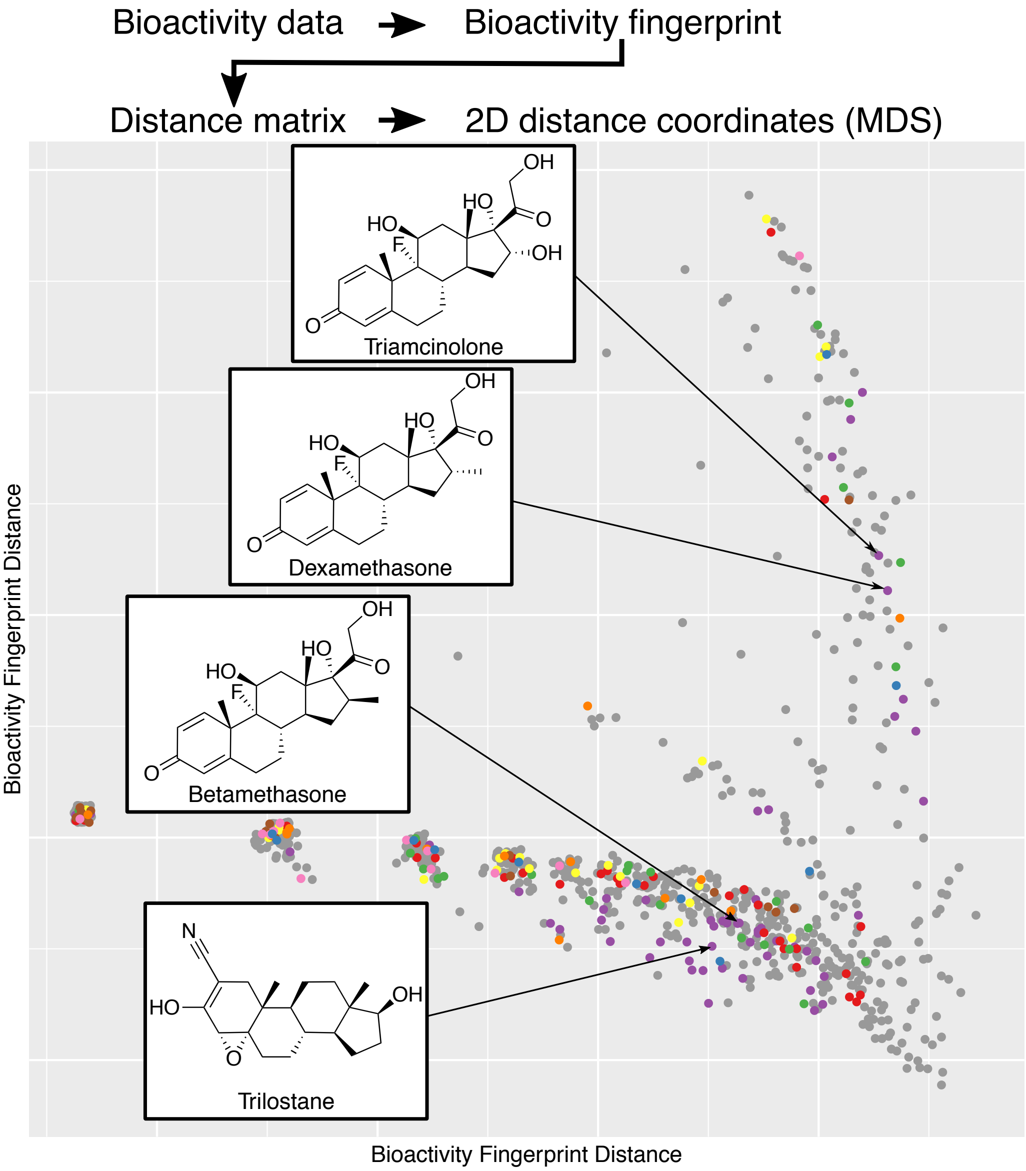


Figure 4.2: **Clustering small molecules by structure and bioactivity profile.** Multidimensional scaling (MDS) was used to embed small molecules into a two dimensional space (x- and y-axis). Each point represents an FDA approved drug. The distance between the points is proportional to the similarity of the bioactivity profile available for each drug. Structure-based clustering results are also indicated by colors for 8 mid-size clusters, while the remaining compounds are shown in grey. The latter clustering used atom pair fingerprints as similarity measure and hierarchical clustering with tree cutting to assign compounds to discrete structure-based similarity groups. Four example compounds are shown, which are members of the same structural cluster (shown here in purple), which is comprised largely of steroid drugs.

4.2 Time performance evaluation

The time performance of bioassayR functions can be divided into two categories: database loading, and database querying. The former is only relevant to users planning to create a custom *bioassayDB* database file. In the case of database loading, the major performance bottleneck involves parsing raw text files into a machine readable format (using regular expressions internally), and therefore is expected to scale in roughly $O(n)$ (linear) time with the number of compound-target activity results, with limited dependence on the absolute number of assays, targets, and compounds. In the case of database querying, the major performance bottleneck involves searching the database file, which makes use of SQLite hash table indices engineered to match their corresponding bioassayR query functions. In most cases these are expected to scale in roughly $O(log\ n)$ (logarithmic) time in proportion to the number of distinct compound-target activity results stored in the database. For repeated queries of the same type (such as finding the active targets for each compound in a large compound library), each query will take a similar amount of time.

Function	Total Compounds				
	1	10	100	1000	10000
Parse assay from raw files (*parsePubChemBioassay*)	0.01	0.01	0.02	0.09	0.86
Loading assay into SQL database (*loadBioassay*)	0.01	0.01	0.01	0.01	0.03
Finding all active protein targets (*activeTargets*)	0.01	0.06	0.52	5.77	60.77
Building *bioassaySet* by cids (*getBioassaySetByCids*)	11.28	18.52	27.14	61.62	144.13
Building HTSFP matrix (*bioactivityFingerprint*)	0.01	0.01	0.05	0.72	11.42
Binary fingerprint search (*fpSim*)	0.00	0.00	0.00	0.02	0.28
Trinary fingerprint search (*trinarySimilarity*)	0.00	0.00	0.02	0.20	2.44

Table 4.4: **BioassayR function runtime.** Total runtime in seconds for seven key bioassayR functions over the range of 1-10,000 randomly selected compounds. Results generated on a single cpu core of an AMD Opteron 6376 2.3 Ghz processor.

Table 4.4 reports the total elapsed runtime for seven key bioassayR functions, across compound libraries of size 1, 10, 100, 1000, and 10000. In order, these functions represent the steps of a possible workflow involving loading a *bioassayDB* database, identifying compounds of interest based on their active targets, and then generating and analyzing two types of HTSFPs based on these compounds. These numbers were generated in a single-threaded manner, using one core of a 16 core AMD Opteron 6376 2.3 Ghz processor on a modern rack-mount server running CentOS Linux. Assay parsing and loading statistics were generated by loading subsets of PubChem Bioassay assay aid 360 into an empty database file. The remaining (query based) benchmarks were generated from the full pre-built PubChem BioAssay database, using a random sample of highly screened (at least 10 distinct target assays) compounds. Binary HTSFP searches were performed using the ChemmineR *fpSim* function [2]. The slowest performing function (*getBioassaySetByCids*) involves translating bioactivity data from the *bioassayDB* database file into an in-memory sparse matrix, and is typically only used for a small subset of the available bioactivity data, relevant to a specific analysis. Despite it's runtime, it can be used for a large number of compounds and targets, as it scales with less than linear $O(n)$ time.

While not necessary for most analyses, users can optionally accelerate a large analysis by running bioassayR in parallel across multiple cpu cores by dividing batch queries into equally sized smaller jobs, and executing in parallel using the R library foreach [1]. To enable this, it is preferable for each thread to open it's own read-only access to the same *bioassayDB* database file, stored on a commonly accessible ramdisk or solid-state hard drive.

4.3 Bibliography

[1] Revolution Analytics and Steve Weston. *foreach: Provides Foreach Looping Construct for R*, 2015.

[2] Y Cao, A Charisi, L C Cheng, T Jiang, and T Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24(15):1733–1734, July 2008.

[3] J Chen, S J Swamidass, Y Dou, J Bruand, and P Baldi. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, 21(22):4133–4139, November 2005.

[4] Margaret J Gabanyi, Paul D Adams, Konstantin Arnold, Lorenza Bordoli, Lester G Carter, Judith Flippen-Andersen, Lida Gifford, Juergen Haas, Andrei Kouranov, William A McLaughlin, David I Micallef, Wladek Minor, Raship Shah, Torsten Schwede, Yi-Ping Tao, John D Westbrook, Matthew Zimmerman, and Helen M Berman. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *Journal of Structural and Functional Genomics*, 12(2):45–54, April 2011.

[5] A Gaulton, L J Bellis, A P Bento, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, D Michalovich, B Al-Lazikani, and J P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(D1):D1100–D1107, December 2011.

[6] T Girke. ChemMine. A Compound Mining Database for Chemical Genomics. *Plant Physiol.*, 138(2):573–577, May 2005.

[7] J C Gower. Multivariate Analysis and Multidimensional Geometry. *The Statistician*, 17(1):13, 1967.

[8] E A Howe, A de Souza, D L Lahr, S Chatwin, P Montgomery, B R Alexander, D-T Nguyen, Y Cruz, D A Stonich, G Walzer, J T Rose, S C Picard, Z Liu, J N Rose, X Xiang, J Asiedu, D Durkin, J Levine, J J Yang, S C Schurer, J C Braisted, N Southall, M R Southern, T D Y Chung, S Brudz, C Tanega, S L Schreiber, J A Bittker, R Guha, and P A Clemons. BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.*, 43(D1):D1163–D1170, January 2015.

[9] John J Irwin and Brian K Shoichet. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45(1):177–182, January 2005.

[10] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.*, 52(7):1757–1768, July 2012.

[11] T Liu, Y Lin, X Wen, R N Jorissen, and M K Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, 35(Database):D198–D201, January 2007.

[12] L Matthews, G Gopinath, M Gillespie, M Caudy, D Croft, B de Bono, P Garapati, J Hemish, H Hermjakob, B Jassal, A Kanapin, S Lewis, S Mahajan, B May, E Schmidt, I Vastrik, G Wu, E Birney, L Stein, and P D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37(Database):D619–D622, January 2009.

[13] Sereina Riniker, Yuan Wang, Jeremy L Jenkins, and Gregory A Landrum. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.*, 54(7):1880–1891, July 2014.

[14] K P Seiler, G A George, M P Happ, N E Bodycombe, H A Carrinski, S Norton, S Brudz, J P Sullivan, J Muhlich, M Serrano, P Ferraiolo, N J Tolliday, S L Schreiber, and P A Clemons. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, 36(Database):D351–D359, December 2007.

[15] Uma T Shankavaram, Sudhir Varma, David Kane, Margot Sunshine, Krishna K Chary, William C Reinhold, Yves Pommier, and John N Weinstein. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*, 10(1):277, 2009.

[16] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 43(D1):D1049–D1056, January 2015.

[17] Y Wang, J Xiao, T O Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, B A Shoemaker, E Bolton, A Gindulyte, and S H Bryant. PubChem's BioAssay Database. *Nucleic Acids Res.*, 40(D1):D400–D412, December 2011.

[18] Yanli Wang, Tugba Suzek, Jian Zhang, Jiyao Wang, Siqian He, Tiejun Cheng, Benjamin A Shoemaker, Asta Gindulyte, and Stephen H Bryant. PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, 42(D1):D1075–D1082, December 2013.

[19] D S Wishart, C Knox, A C Guo, D Cheng, S Shrivastava, D Tzur, B Gautam, and M Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database):D901–D906, December 2007.

Chapter 5

Large-scale Bioactivity Analysis of the Druggable Proteome

5.1 Abstract

This study presents a large-scale analysis of the small molecule bioactivity profiles across large quantities of diverse protein families represented in PubChem BioAssay screening data. We compare the bioactivity profiles of FDA approved drugs to non-FDA (other) compounds, and report several distinct patterns characteristic of the approved drugs. We found that a large fraction of the previously reported higher target promiscuity among FDA approved compounds, compared to non-FDA bioactives, is due to cross-reactivity within rather than across protein families. This trend is not the result of more extensive assay testing of certain target classes. We identified 804 potentially novel protein target candidates for FDA approved drugs, as well as 901 potentially novel target candidates with active non-FDA compounds, but no active FDA approved drugs. We also identified 486348 potentially novel compounds active against the same targets as FDA approved drugs, as well as 153402 potentially novel compounds active against targets without active FDA approved drugs. By quantifying the agreement among replicated screens, we estimate that more than half of these novel outcomes are reproducible and likely to be correct. We used biclustering to identify clusters of FDA approved drugs with enriched activity against a common set of protein targets. We also infer compound promiscuity with a Bayesian statistical model, and assess the sensitivity and specificity of two common methods for identifying promiscuous compounds. Aggregator assays exhibit greater accuracy in identifying highly promiscuous compounds, while PAINS substructures are able to identify a much larger set of “middle range” promiscuous compounds. Additionally, we report a large number of promiscuous compounds not identified as aggregators or PAINS. In summary, the results of this study represent a rich reference for selecting novel drug and target protein candidates, as well as for eliminating candidate compounds with unselective activities.

5.2 Introduction

High throughput screening (HTS) is a key technology for identifying bioactive small molecules for chemical genomics and drug discovery applications. The discovery of small molecules exhibiting a high level of efficacy and specificity is confounded by many challenges, including experimental noise in HTS experiments, and an extremely large search space. The potentially testable compound-protein target space consists of nearly two trillion possible combinations, if we regard each of the over 91 million small-molecules in the PubChem compounds database (at the time of writing) as potential drug candidate, and each of the annotated protein coding genes in the *Homo sapiens* genome (19950 genes according to GENCODE 25) as a potential drug target [27, 16]. This search space becomes much larger if we consider alternative splicing, non-protein biomolecule targets, and potential targets from other species, *e.g.* microbiome targets and parasite targets. In recent years, a substantial number of small molecule vs protein target assays have become available in the public domain, which investigate a portion of this search space. At the time of writing, the PubChem BioAssay database contains just over 230 million small molecule bioactivity outcomes, over half of which involve activity against a clearly defined protein target [50]. It includes most of the bioactivity data available in the public domain as it imports assays from many sources such as ChEMBL, and also provides negative (inactive) assay outcomes not reported in many databases[13]. This large data volume presents an opportunity to systematically investigate small molecule-target interactions, with the potential to provide insights relevant to future drug discovery efforts [25, 53, 51, 15, 5, 49, 50]. These data also have potential utility for identifying and excluding drug candidates with undesirable binding properties (*e.g.* unselective promiscuous binders), developing multi-target (polypharmacological) drug treatments, predicting potential side and toxic effects of small molecules, and assessing the druggability of novel target proteins [15, 36, 43, 29, 34, 21, 38, 14, 30, 18, 9]. The following gives a brief overview of previous work in this field.

Shortly after the NIH Molecular Library Roadmap Initiative made available large public screening data in PubChem BioAssay, Han *et al.* reported the distribution of assay participation, target selectivities, and target diversity in these data, while Zhang *et al.* later reported bias in target and compound selection among these data [55, 15, 56]. Hu and Bajorath quantified the distribution of active target proteins in the PubChem, DrugBank, and ChEMBL databases, and found that 37.4% of FDA approved drugs interact with more than five targets, while other active compounds tend to interact with only 1-2 targets, with only a 7.6% probability of more than five [20, 52, 13]. Recently, Jasial *et al.* analyzed compound promiscuity in PubChem BioAssay and found a median of 2 active targets for non-FDA approved compounds [24]. In comparison to previous work in this field, our study is unprecedented by providing a broad in-depth analysis of the publicly available small molecule bioactivity space, including target selectivity profiles within and across protein families considering variable evolutionary distances.

The concept of target selectivity has been introduced in previous literature, in order to quantify the number of distinct protein targets a compound exhibits activity against. Two common metrics for quantifying target selectivity have been frequently used. First, the total number of active targets across all participating assays, is referred to simply as target selectivity [15]. Second, the fraction of actives out of the total number of screened targets has been referred to as the hit ratio [8]. We analyze target selectivity with both methods, as they have complementary strengths and weaknesses.

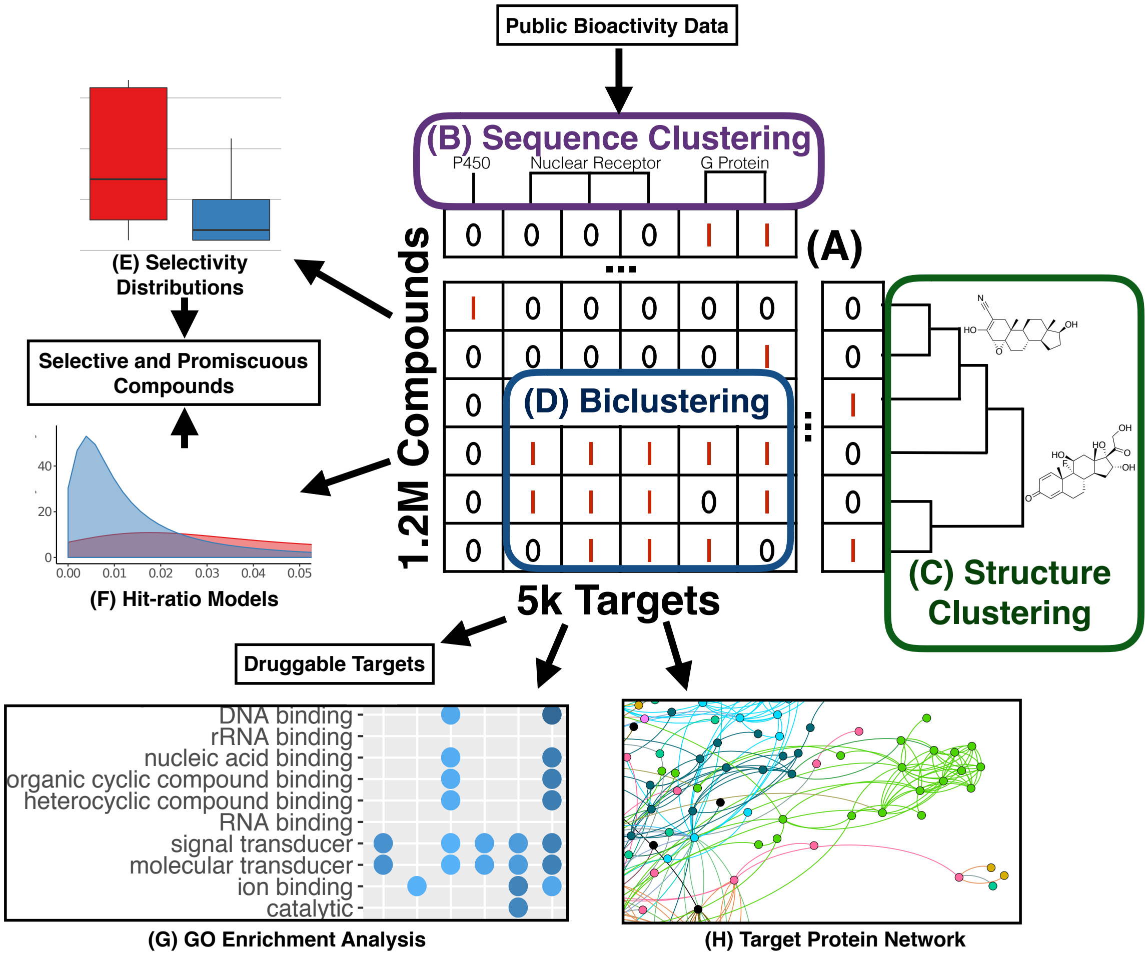


Figure 5.1: **Bioactivity data mining strategy.** Public bioactivity data was first summarized in a compound-target bioactivity matrix (**A**). Protein targets and small molecules were clustered by sequence (**B**) and structure (**C**) respectively, and compound-target sets with shared bioactivity profiles were identified with biclustering (**D**). For small molecules, the distributions of (**E**) target selectivity (the number of active targets) and (**F**) hit ratio (the fraction of screened targets that are active) were quantified. For protein targets, enriched GO (Gene Ontology) terms (**G**) among proteins with common bioactivity were identified, and a network (**H**) was constructed which connects target proteins with similar bioactivity profiles. These analyses highlight several interesting bioactivity patterns, identify promiscuous and selective compounds, and identify druggable protein targets and protein domains.

In this study, we mine large public bioactivity data to investigate many outstanding questions about the patterns of target selectivity among small molecules. Fig 5.1 provides a visual overview of important steps in our data analysis strategy. To investigate why FDA approved drugs on average exhibit activity against a greater number of targets than non-FDA compounds, we computed the target selectivity of small molecules against protein clusters obtained with three distinct methods that classify protein sequences across increasingly large evolutionary distances. While the FDA approved drugs have on average a greater number of targets, these targets more frequently share sequence similarity than targets of non-FDA active compounds. We also found that many of these multi-target FDA approved drugs fall into biclusters, where a common set of drugs share activity against a common set of protein targets that are enriched for common molecular function annotations, suggesting a shared chemical mechanism leading to cross-reactivity. To determine which targets are more accessible to small molecule perturbations than others, we quantified the number of active compounds for targets grouped by shared protein domains, and found active compounds for targets representing 32.4% of the domains present in the *H. sapiens* proteome. Clustering the targets by similar amino acid sequences, we found 9120 active target outcomes for FDA approved drugs not currently annotated in drug databases. By quantifying the rate of agreement among millions of replicated compound-target pairs across distinct assays, we estimate that over half of these novel results are accurate bioactivity outcomes. To investigate the frequency of highly promiscuous compounds, we used a statistical model to infer the hit ratio of each compound, and report 1157 likely-promiscuous compounds not previously identified by two common methods of identifying promiscuous compounds, aggregator assays and PAINS substructures [36, 2].

5.3 Results and Discussion

5.3.1 Bioactivity Data

Bioactivity Data Curation and Overview

The bioactivity data analyzed by this study were downloaded from PubChem BioAssay on April 6th, 2016. They included 1.2 million distinct small molecule structures tested against 5069 protein targets in 68029 assay experiments [50]. We were able to utilize all experiments annotated with a single clearly defined protein target, and reporting an active score for at least one small molecule. Assays with no active scores, or no machine readable protein target annotation were excluded. Much of this data summarizes the results from primary screening experiments which provide only binary active/inactive results, but we also analyze confirmatory assays, if binary calls are also provided.

As compounds were screened against variable numbers of targets, the compound vs target bioactivity space obtained from PubChem BioAsssay is sparse. Currently, there are 162 million compound-target activity records available, populating 2.6% of the full bioassay matrix with at least one measurement. Within the explored bioactivity space, active values are relatively rare (just over 2.3 million), representing just over 1.3% of total tested values, or about 0.027% of the total space. If we consider just the

566983 “highly screened” compounds tested against at least 10 distinct targets, the density of tested bioactivity outcomes increases to 6.1%. The patterns of bioactivity among these “highly screened” compounds are the focus of this study, as they provide information about bioactivity profiles across many targets. Collapsing the protein target space by merging very similar sequences such as truncations, close orthologues, and paralogues reduces these targets from 5069 protein targets to 2249 target clusters, producing a smaller and more dense bioactivity matrix. This is described in more detail in the Methods section (see “Clustering Protein Targets by Sequence”). A subset of the bioactivity space is non-sparse, with a set of 81660 compounds by 247 target clusters that has been explored 100%, which we discuss in Supporting Information and provide as a downloadable reference for users in S7 File.

To facilitate comparisons throughout this study among FDA and all other compounds, we obtained a list of the 1173 FDA approved drugs with known PubChem ids from the DrugBank database (version 4.2), and quantified the number of screened targets for both categories[52]. Table 5.1 shows the distribution of total screened protein targets for the compounds in PubChem BioAssay. The overall distribution is also plotted in Fig 6.1 in Supporting Information. While a disproportionately large fraction of non-FDA compounds were screened against a small number of targets, the distribution of screening frequencies between highly screened FDA approved and non-FDA compounds is similar. Highly screened FDA approved drugs were screened against a mean of 242 targets (median 184), while highly screened non-FDA compounds were screened against a mean of 224 (median 280) targets. Therefore, these data allow us to compare patterns of target selectivity between many FDA approved and non-FDA compounds with similar assay participation profiles. Additionally, the hit ratio statistical model we introduce below in the “Promiscuous Binders and Hit Ratio Statistical Model” section provides a robust method of comparing target selectivity across compound sets with varying assay participation.

Screened Protein Targets	FDA Approved Drugs	Non-FDA Compounds
1	31	359135
2-4	41	135786
5-9	27	151385
10-49	197	150202
50-99	128	51849
100-199	150	30277
200-299	85	69098
300-399	94	202225
400-499	106	63219
≥500	139	82

Table 5.1: **Screening frequency of FDA approved and non-FDA compounds against increasing numbers of protein targets.** Data is included from all assay experiments in PubChem BioAssay annotated with one clearly defined protein target, and reporting an active score for at least one small molecule. Multiple assays against the same target are counted only once.

When comparing the bioactivity profile and target selectivity among compounds, we focus on compounds with evidence of activity against at least one protein target, as active compounds are more likely to be of biological or medical interest. Of the 566983 “highly screened” compounds mentioned above, 312308 have also been found active against one or more targets. Among the highly screened active compounds, 759 are FDA approved drugs, whose patterns of target selectivity we compare and contrast with non-FDA highly screened active compounds.

Data Quality and Reproducibility

Systematically analyzing public bioactivity data presents many data quality challenges stemming from experimental error, and missing or incorrect annotation. While efforts such as the BioAssay Research Database (BARD) and BioAssay Ontology are underway to curate a set of assays with detailed high quality annotations, these represent a very small subset of the publicly available bioactivity data [48, 19]. To assess the reliability of the data, we estimate an error rate for compound-target combinations tested multiple times in separate assays by quantifying how often the results agree or disagree. This estimate combines all errors causing *in vitro* screening outcomes from different primary screening assays to disagree, such as underlying experimental noise, data curation and annotation errors, as well as disagreement resulting from unique experimental context or conditions for a particular assay, that are not provided in a machine readable format. While we can quantify how often activity outcomes disagree across different assays, the PubChem BioAssay data does not include information about the exact cause of a disagreeing activity outcome, or in which assay the error occurred in. As this data includes assays of variable design and robustness, individual assays will have different error rates. However, our estimate represents the probability of any individual compound-target activity outcome reporting an incorrect result when combined and analyzed in aggregate, as we do in this study.

Table 5.2 shows the number of distinct compound-target pairs that were screened a given number of times. In Supporting Information we solve algebraically for the error rate based on the frequency of agreeing or disagreeing sets among compound-target pairs tested in exactly two different assays. As explained and justified in Supporting Information, our estimate is an approximation which requires two simplifying assumptions. First, we set the average false positive and false negative rates across the entire dataset equal, and estimate an overall error rate e. Second, we assume that the fraction of true active compound-target pairs in the total PubChem BioAssay data is approximately the same for both the set with two replicates, and the larger set of data with more or less than two replicates. Based on these assumptions, we estimate an error rate of approximately 0.698%, representing the probability of any individual bioactivity outcome reporting the opposite of its true result.

Times Screened	Number of Compound-Target Pairs
2	21220270
3	3308744
4	726700
5	29610
≥6	376787

Table 5.2: **Screening Frequency.** The number of distinct compound-target combinations screened in multiple assays, listed for increasing numbers of assays.

The high throughput screening experiments we analyze here require choosing a hit threshold, which assigns a binary active or inactive outcome to each compound tested, based on the magnitude of its experimentally measured activity level. The specific hit threshold is a subjective choice of the experimentalist that balances the acceptability of false positives and false negatives for a given purpose, and is not provided to us in a machine readable manner. Thus, it is not feasible to provide here a precise estimate of the fraction of actives which are true positives [35]. In many drug discovery efforts, false positives are more problematic than false negatives. As a result, experimentalists are more likely to chose a stringent activity cutoff, biased towards avoiding false positives. As such, these data suggest a rough estimate on the lower bound of the fraction of active compound-target outcomes which are true positives of approximately 66%, but it may be higher. Therefore, we expect that despite a considerable error rate, more than half of the unreplicated positive activity outcomes in these data are meaningful in the context they are used for in this study.

Protein Target Diversity

To assess the target protein diversity represented in PubChem BioAssay, we enumerated the number of distinct targets by three methods which group targets across increasingly large evolutionary distances, including (i) unique protein identifiers, as well as clustering (ii) by protein sequence similarity and (iii) by Pfam domains [41]. The 68029 assay experiments we analyzed grouped into 5069 clusters of assays sharing an identical distinct GenBank Protein GI (Gene Identifier), each of which has a unique amino acid sequence [40]. By clustering these targets together based on an amino acid sequence identity of at least 60%, an E-value $\leq 10^{-4}$, and an alignment coverage of at least 80% for the longer sequence, we identified 2249 distinct target clusters. This method clusters together minor truncations engineered for screening purposes, close orthologues, and close paralogues (see "Clustering Protein Targets by Sequence" in Methods).

To investigate target diversity at the domain level, we mapped Pfam-A domains to each protein target with a distinct GI as described in the "Protein Annotations and GO Enrichment" section of Methods. We quantify the distribution of screening participation by active and inactive compounds for the targets with different Pfam domains in Fig 5.2, with all domains shown in the left panel, and the subset present in the *H. sapiens* proteome shown in the right panel. In total there are 2838 distinct Pfam domains represented in PubChem BioAssay that are associated with active compounds, and therefore have evidence of druggability. In comparison, 32.4% of the 7431 distinct Pfam domains represented in the *H. sapiens* proteome are also represented in these bioassays and report active compounds, while 27.1% of them were screened directly on *H. sapiens* proteins. Additionally, there are proteins with active compounds in PubChem BioAssay which contain 433 Pfam domains not present in the *H. sapiens* proteome, many of which are domains restricted to bacteria and plants. There are 795 Pfam domains that are extremely highly screened, with activity outcomes for over 200k compounds each. The number of domains with active compounds is greater than inactive compounds, due to assays which do not report inactive outcomes. While these are mostly small assays reporting few activity outcomes, they substantially increase the information about the druggable space by reporting active compounds for 792 domains not present in the other assays, 614 of which are present in the human proteome. Because proteins often contain multiple Pfam domains, duplications are unavoidable with this type of protein family clustering. Nevertheless, the results reported here are a reasonable approximation of the diversity of protein families represented in PubChem BioAssay.

Fig 5.3A enumerates the relative abundance of active FDA approved drugs, active non-FDA compounds, and total protein targets for the 35 domains with the largest

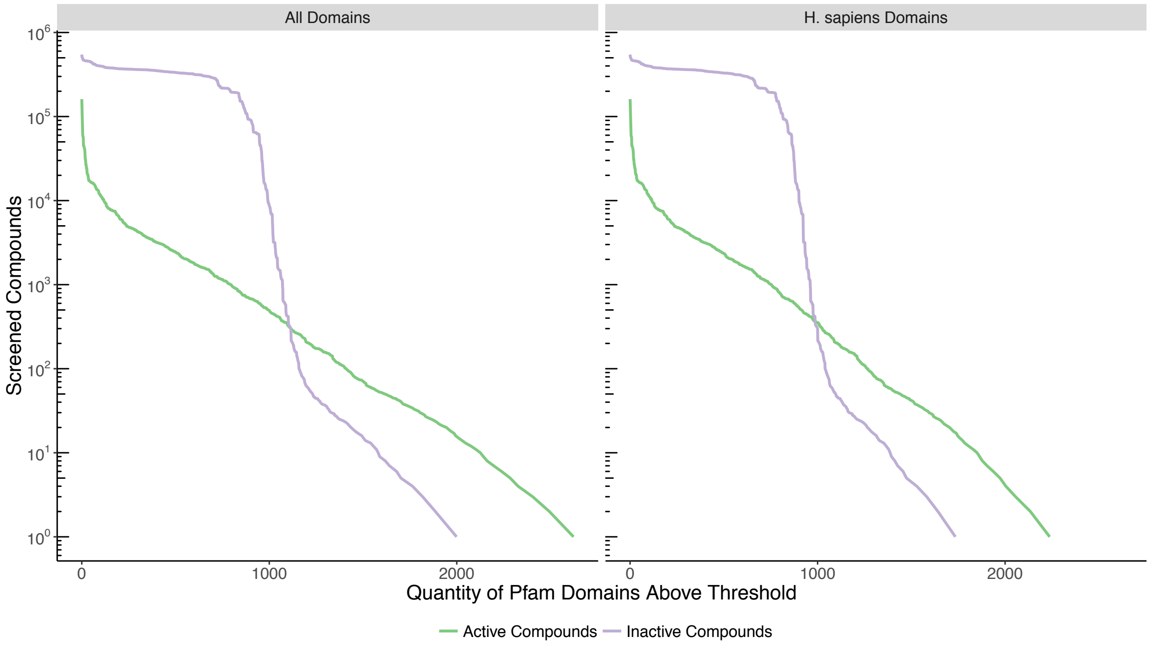


Figure 5.2: **Pfam domain screening participation.** The quantity of Pfam domains is plotted on the horizontal (x) axis whose protein targets have at least as many active or inactive compounds as shown on the vertical (y) axis. The left panel includes all Pfam domains in the PubChem BioAssay targets, while the right panel includes just those domains also present in the *H. sapiens* proteome, including non-*H. sapiens* targets which share a common domain with an *H. sapiens* protein.

number of active FDA approved drugs. As can be seen by comparing column 1 (FDA Approved Drugs) to column 2 (Non-FDA Compounds), the fraction of screened compounds active against each domain is significantly higher for FDA approved drugs. However, due to the much greater number of non-FDA compounds, the total number of non-FDA actives is much higher than the number of FDA approved drugs. For comparison, the number of proteins within each Pfam domain is given in Column 3 of Fig 5.3A for both proteins represented in PubChem BioAssay and those encoded in the *H. sapiens* genome. In some cases, the number of targets in PubChem BioAssay exceeds those in the *H. sapiens* proteome because it includes targets from other species, as well as engineered targets (*e.g.* truncations) developed for screening purposes. The proteins targeted by the greatest number of FDA approved drugs include rhodopsin-like GPCRs, cytochrome P450 enzymes, and nuclear hormone receptors, with a large number of non-FDA compounds also targeting these proteins.

Several domains such as Tyrosyl-DNA phosphodiesterase and protein kinase have a large number of active non-FDA compounds compared to the ordering of domains, which is based on a decreasing number of FDA approved active compounds. In cases such as kinases this may be explained by the large number of assays screening distinct targets within a large protein family. However, in other cases such as Tyrosyl-DNA phosphodiesterases there are a comparatively large number of active non-FDA compounds despite relatively fewer target proteins screened within this much smaller family. Overall, we found 486348 non-FDA compounds active against individual tar-

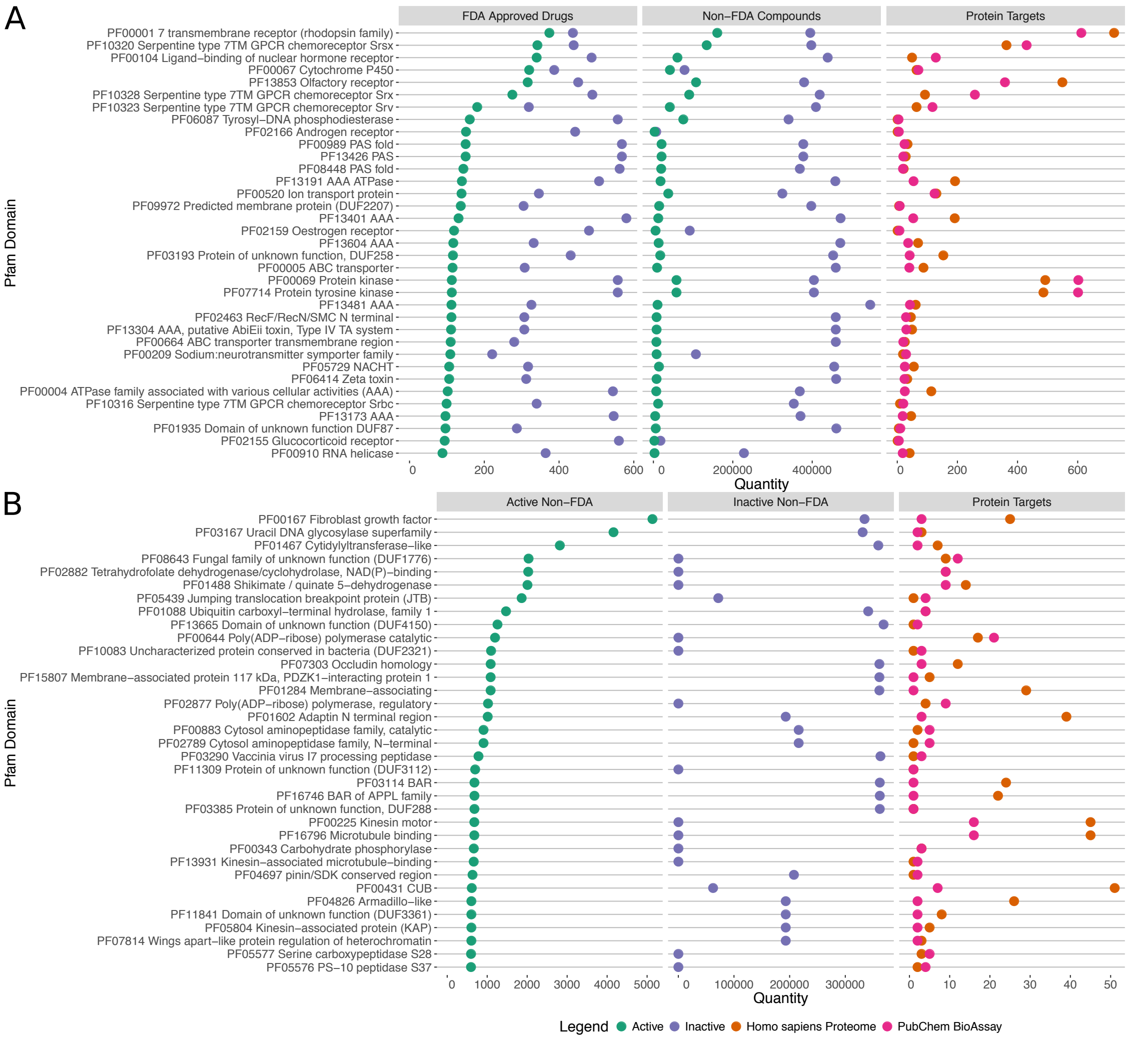


Figure 5.3: **Frequency of active PubChem BioAssay compounds across protein target domains.** The target proteins represented in PubChem BioAssay have been classified by Pfam protein domains present in the *H. sapiens* proteome (vertical axis). We report data for all proteins which encode a Pfam domain present in the *H. sapiens* proteome, even if the assay was performed against a protein from another species. We show here only domains with at least 100 amino acid residues in the homology model, to avoid small repeats and domains unlikely to be drug targets. The quantity of targets with each domain among the PubChem BioAssay data, and within the *H. sapiens* proteome (all proteins, including those not screened in PubChem BioAssay) are shown on the right in both plots. **(A)** The top 35 Pfam domains with the greatest number of active FDA approved drugs, in decreasing order. **(B)** The top 35 Pfam domains with the greatest number of non-FDA compounds, but no active FDA approved drugs, in decreasing order. A full table with the number of active compounds for each domain, including non-*H. sapiens* domains, and domains with under 100 residues is provided in the S2 File of Supporting Information.

gets (distinct GenBank GIs) that also have active FDA approved drugs, representing a set of potentially novel compounds active against potentially therapeutic targets.

As we reported previously, the FDA approved drugs show activity against 1789 protein targets, whereas the non-drug compounds show activity against an additional 3020 protein targets, of which 901 are substantially distinct at the sequence level, based on the clustering by sequence similarity mentioned above [1]. Fig 5.3B lists the top 35 domains with no active FDA approved compounds, but the greatest number of non-FDA compounds. These targets with domains not known to be accessible to FDA approved drugs represent a greatly expanded space of potentially druggable targets and small molecule drug candidates. In total, we found 153402 compounds active against individual targets (distinct GenBank GIs) with no active FDA approved drugs. While some of these compounds will be false positives due to experimental noise, the magnitude of actives suggests a large quantity of truly active compounds. We provide the number of active FDA approved and non-FDA compounds for the full set of Pfam domains in the S2 File of Supporting Information.

5.3.2 Target Selectivity

Target Selectivity Distribution

Highly screened bioactive small molecules can be categorized according to target selectivity, which is the number of distinct protein targets they show activity against. By quantifying the distribution of target selectivities, we can identify highly selective and less selective compounds, as well as compare the selectivities of FDA approved drugs to non-FDA compounds. To address this, we computed the distribution of target selectivities among the highly screened active compounds in PubChem BioAssay, each of which were tested against 10 or more protein targets, and active against at least one. We computed target selectivity based on the three types of protein clustering methods mentioned in the previous section. “Target selectivity” counts each target with a distinct amino acid sequence (distinct GenBank Protein Gene Identifiers) separately, while “cluster selectivity” counts the number of sequence-based clusters a compound shows activity against. Third, “domain selectivity” counts activity against any set of targets sharing a common Pfam protein domain only once. Due to the existence of protein targets with multiple domains, we compute the number of domain clusters independently for each compound. For example, if a compound is active against 5 targets, but 4 share a common domain, its domain selectivity is 2. This is the same as counting the number of connected components in a graph where each node represents an active protein target, and edges represent target pairs sharing a common Pfam domain. The distribution of counts for all three clustering methods is shown in Fig 5.4. Fig 5.4B includes a boxplot which highlights the quantiles for each distribution, while in Fig 5.4A, counts are shown for values up to 20. There are an additional 144 FDA approved drugs and 6285 non-FDA compounds with greater than 20 individual active targets. These are not shown in Fig 5.4A as they represent a very small fraction of the total compounds, and for highly promiscuous compounds, may simply represent the number of screened targets instead of a biologically relevant quantity. The “Promiscuous Binders and Hit Ratio Statistical Model” section below quantifies and visualizes selectivity in a way that normalizes by assay participation, allowing us to investigate selectivity distributions among highly promiscuous compounds.

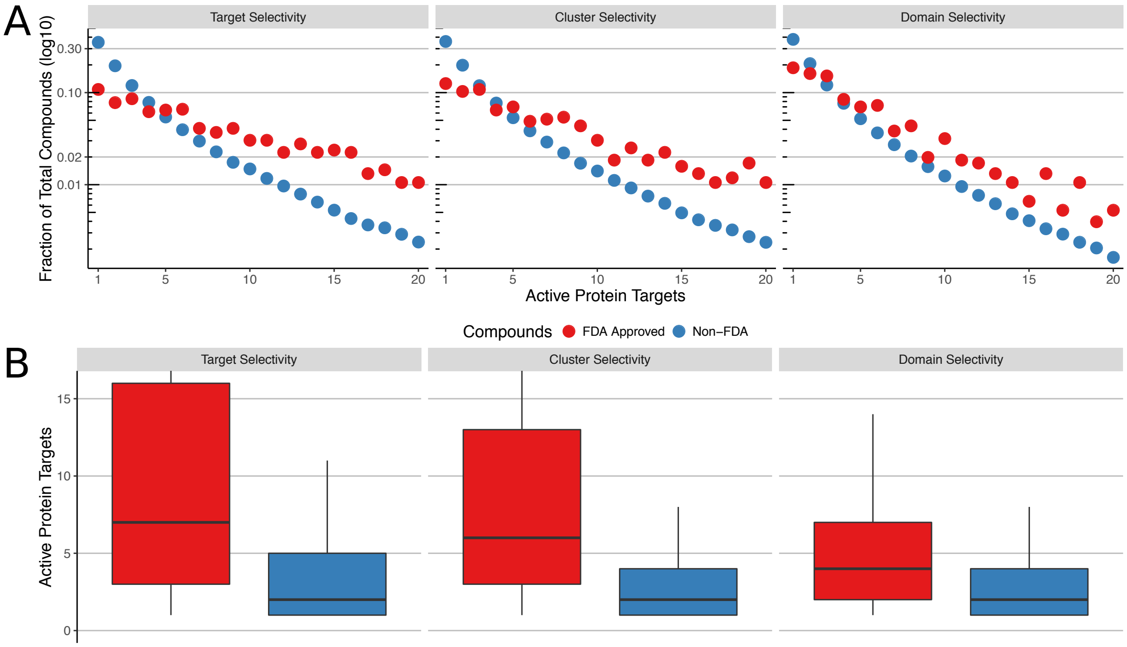


Figure 5.4: **Distribution of active proteins per compound.** Both plots show the distribution of target selectivity among PubChem BioAssay compounds, with selectivity quantified by three methods which cluster protein targets across increasingly large evolutionary distances, as described in the text. FDA approved drugs (red) are shown separately from non-FDA compounds (blue). **(A)** Semi-log plot of the target selectivity distributions, where horizontal (x) axis represents the number of active protein targets and/or protein target clusters, while the vertical (y) axis represents the fraction of each compound set that is active against a particular number of targets. **(B)** Box plot of the target selectivity distributions, with horizontal lines at the 25%, 50%, and 75% quantiles for each distribution. The vertical (y) axis represents the number of active protein targets and/or protein target clusters. Whiskers extend to 1.5 times the inter-quartile range, however we limit the vertical (y) axis to 16 in order to zoom into the higher density region.

Interestingly, the FDA approved drugs show a much greater frequency of activity against many targets, and reduced frequency of activity against only one or two targets as compared to non-FDA compounds, as shown in Fig 5.4. In Supporting

Information we also provide a table with median, mean, and trimmed mean values for all three clustering methods. We performed a one sided Mann-Whitney-Wilcoxon test to determine if the FDA approved drugs have higher counts than the non-FDA compounds vs. the null hypothesis that they have equal or lower counts. The values W were 180M, 173M, and 154M for target, cluster, and domain selectivity counts respectively, with p-value $< 2.2 * 10^{-16}$ by normal approximation for all three comparisons. While this observation was reported in previous literature, we report an even higher number of targets for the FDA approved compounds, based on the larger volume of data we analyze here[23, 20]. This higher number of active targets is unlikely to be due to biased assay participation, because as discussed in the above "Bioactivity Data Curation and Overview" section, non-FDA compounds were screened against a higher median number of targets, yet show a lower median number of active targets. Additionally, in the "Promiscuous Binders and Hit Ratio Statistical Model" section below, we report that this trend is still present when analyzed with a statistical model that accounts for the individual assay participation of each compound.

As shown in Fig 5.4, the target promiscuity of FDA approved drugs decreases substantially as related targets are clustered across increasingly large evolutionary distances, while the target promiscuity of non-FDA compounds decreases to a much lesser extent. For the FDA approved drugs, the median selectivity drops from 7 to 4 when targets sharing common domains are clustered. The Mann-Whitney-Wilcoxon test results in the previous paragraph also quantify the magnitude of this difference. As this test is based on position in a ranked list, the result indicates that for a large number of FDA approved compounds, there are a greater number of non-FDA compounds with higher domain selectivity counts, than the number of non-FDA compounds with higher target selectivity counts. These results highlight a fundamental difference in the overall trend of bioactivity between FDA and non-FDA compounds active against many targets. While the FDA approved drugs tend to be active against many more targets than non-FDA compounds, a greater fraction of these targets share common Pfam domains and/or overall sequence similarity. While a substantial fraction of the active targets of FDA approved drugs are closely related, the FDA approved drugs also exhibit activity against a slightly higher number of unrelated targets than do the non-FDA compounds. We further explore the selectivity against distinct Pfam domain families in the next section. As a resource for readers, we report the target selectivity, cluster selectivity, and domain selectivity for all highly screened actives in the S2 File of Supporting Information.

Selectivity Across Pfam Domains

As the FDA approved drugs exhibit wide variation in target selectivity, with both highly selective, and highly promiscuous compounds, we wanted to determine whether promiscuous and selective compounds exhibit activity against different subsets of the protein target space. To answer this question, we identified the highly screened compounds exhibiting activity against the target proteins grouped by Pfam domains. We then computed for each domain the median domain selectivity counts of the active compounds. Domain selectivity is the same as introduced in the "Target Selectivity Distribution" section above, where active targets sharing a common domain are counted only once. We performed this separately for the FDA approved, and non-

FDA compounds, while excluding domains with less than 10 active compounds from both sets. Table 5.3 quantifies the number of domains grouped into 8 bins of median domain selectivity, showing an extremely wide variation of median domain selectivities, including both domains whose active compounds tend to be highly promiscuous, and domains whose active compounds tend to be highly selective. Table 5.4 lists the top 15 Pfam domain families whose active FDA approved drugs are most promiscuous across proteins with different domains, while Table 5.5 lists the top 14 Pfam domain families whose active FDA approved drugs are most selective across proteins with different domains. Domains not present in the *H. sapiens* proteome are not shown in Tables 5.4 and 5.5, but were included in the analysis and are available in the S2 File of Supporting Information.

Bins of Median Domain Selectivity	FDA Approved Domain Counts	Non-FDA Domain Counts
2-4	6	113
5-7.5	80	169
8-10.5	119	81
11-13.5	71	16
14-16.5	63	7
17-19.5	29	3
20-22	17	2
23.5-25.5	6	0

Table 5.3: **Frequency of Pfam domains binned by median domain selectivity of active compounds.** Each row represents a set of Pfam domains whose active compounds (against targets with that domain) have a median domain selectivity in the range specified. Domain selectivity is the same as introduced in the "Target Selectivity Distribution" section above, where active targets sharing a common domain are counted only once. The ranges are ordered from top to bottom by increasing number of distinct domain active targets. We report bin counts separately for FDA Approved and Non-FDA compounds.

Domain	Median Target Selectivity FDA Approved/ non-FDA	Median Cluster Selectivity FDA Approved/ non-FDA	Median Domain Selectivity FDA Approved/ non-FDA
PF11956 Ankyrin-G binding motif of KCNQ2-3	47/7	43/7	25.5/5
PF16178 Dimerisation domain of Ca+-activated chloride-channel, anoctamin	41/8	36/7	25/6
PF03520 KCNQ voltage-gated potassium channel	44.5/7	40/7	25/6
PF04547 Calcium-activated chloride channel	38/8	35.5/7	24/6
PF10488 Phosphatase-1 catalytic subunit binding region	49/14	46.5/14	23.5/12
PF10401 Interferon-regulatory factor 3	46.5/8	45.5/8	22/7.5
PF03165 MH1 domain	46.5/8	45.5/8	22/8
PF03166 MH2 domain	46.5/8	45.5/8	22/8
PF09038 Tumour suppressor p53-binding protein-1 Tudor	43.5/6	36.5/6	21/6
PF13520 Amino acid permease	33/7	31/7	21/6
PF15057 Domain of unknown function (DUF4537)	43.5/7	36.5/6	21/6
PF02518 Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	36/12	31/12	21/10
PF13589 Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	50/12	44/12	21/10
PF00183 Hsp90 protein	50/12	44/12	21/10
PF08605 Fungal Rad9-like Rad53-binding	43.5/7	36.5/6	21/6

Table 5.4: **Top 15 Pfam domains with least selective active drugs.** Only domains present in the *H. sapiens* proteome are shown, and are sorted and selected by decreasing domain selectivity among FDA approved drugs. We also excluded domains with under 100 amino acid residues in the homology model, to avoid small repeats and domains unlikely to be drug targets. To the left of the slash in each column is the median target selectivity of FDA approved compounds active against this domain, and to the right is the median selectivity of non-FDA compounds. Selectivity is quantified by three methods of increasingly grouped protein targets as described in the text. For example, the non-FDA approved compounds active against targets with the PF00183 domain are active against a median of 12 targets with distinct GenBank identifiers, so this domain has a 12 after the slash in the first column.

To determine whether functional activities are enriched within the individual selectivity bins of Table 5.3, we used the Molecular Function Gene Ontology annotations (MF GO) of the corresponding Pfam domains to perform a GO term enrichment test based on the hypergeometric distribution (see “Protein Annotations and GO Enrichment” Methods section) [45]. Since we were mostly interested in enrichments within general functional categories, we restricted this analysis to the slim terms of the MF GO. This allows us to identify functional categories that are more abundant within the selectivity bins than one would expect by chance. Fig 5.5 shows the enriched GO terms (p-value < 0.05) for each bin of target selectivity. We also show in the right

Domain	Median Target Selectivity FDA Approved/ non-FDA	Median Cluster Selectivity FDA Approved/ non-FDA	Median Domain Selectivity FDA Approved/ non-FDA
PF06512 Sodium ion transport-associated	7/4	5.5/4	3/2.5
PF11933 Cytoplasmic domain of voltage-gated Na+ ion channel	8/6	7/6	3/3
PF00144 Beta-lactamase	11.5/6	5.5/6	4/5
PF14580 Leucine-rich repeat	14/6	12/6	4/5
PF00324 Amino acid permease	14/7	7/7	5/6
PF00194 Eukaryotic-type carbonic anhydrase	18/11	12/6	5/2
PF01384 Phosphate transporter family	10/5	8/5	5/5
PF00787 PX domain	21/7	15/7	5/6
PF03522 Solute carrier family 12	13/7	7/7	5/6
PF01593 Flavin containing amine oxidoreductase	13.5/9	10/9	5.5/7
PF00135 Carboxylesterase family	10.5/6	8.5/5	5.5/4
PF02931 Neurotransmitter-gated ion-channel ligand binding domain	13/5	10/4	5.5/2
PF02932 Neurotransmitter-gated ion-channel transmembrane region	13/5	10/4	5.5/2
PF00078 Reverse transcriptase (RNA-dependent DNA polymerase)	10.5/8	8.5/8	5.5/7

Table 5.5: **Top 14 Pfam domains with most selective active drugs.** Only domains present in the *H. sapiens* proteome are shown, and are sorted and selected by increasing domain selectivity among FDA approved drugs. We also excluded domains with under 100 amino acid residues in the homology model, to avoid small repeats and domains unlikely to be drug targets. To the left of the slash in each column is the median target selectivity of FDA approved compounds active against this domain, and to the right is the median selectivity of non-FDA compounds. Selectivity is quantified by three methods of increasingly grouped protein targets as described in the text. For example, the non-FDA approved compounds active against targets with the PF00144 domain are active against a median of 6 targets with distinct GenBank identifiers, so this domain has a 6 after the slash in the first column.

column the total number of protein targets in PubChem BioAssay annotated with each term.

Several target selectivity bins are enriched with a characteristic set of MF GO terms. For example, FDA approved drugs active against oxidoreductase targets appear in a promiscuous bin (14-16.5), whereas drugs targeting binding proteins appear in a more selective bin (5-7.5). Overall, this result demonstrates that the different protein domains represented in PubChem BioAssay can be grouped into those druggable primarily with selective compounds, and those druggable primarily with promiscuous compounds. Interestingly, the patterns of term enrichment are different between the FDA approved and non-FDA compounds, with many of the top target classes druggable by more promiscuous FDA-approved compounds having primarily selective non-FDA active compounds and vice versa. This raises the question of if the selectivity levels characteristic of FDA approved drugs are a necessary property for those compounds therapeutic efficacy, or if more selective non-FDA compounds may also include viable drug candidates with a reduced chance of off-target effects. The highly enriched molecular function terms in some bins also raises the question of if these compound-target interactions may share a characteristic selectivity due to a shared chemical mechanism of bioactivity.

Stretched Exponential Distribution

The distribution of active targets for non-FDA compounds shown in Fig 5.4A show a very regular pattern, with a slight curvature in semi-log space. We found that this distribution is well described by the stretched exponential function shown in Eq 6.6 ($R^2 = 0.99912$ for non-FDA cluster selectivity), including two fit parameters c and x_0. This is not due to the distribution of assay participation, as assay participation has a very irregular pattern with a large number of compounds screened against many targets as shown in Table 5.1 and Fig 6.1 in Supporting Information. Stretched exponential functions are commonly observed in natural multiplicative processes, and we report detailed methods and related citations for this observation in Supporting Information[31].

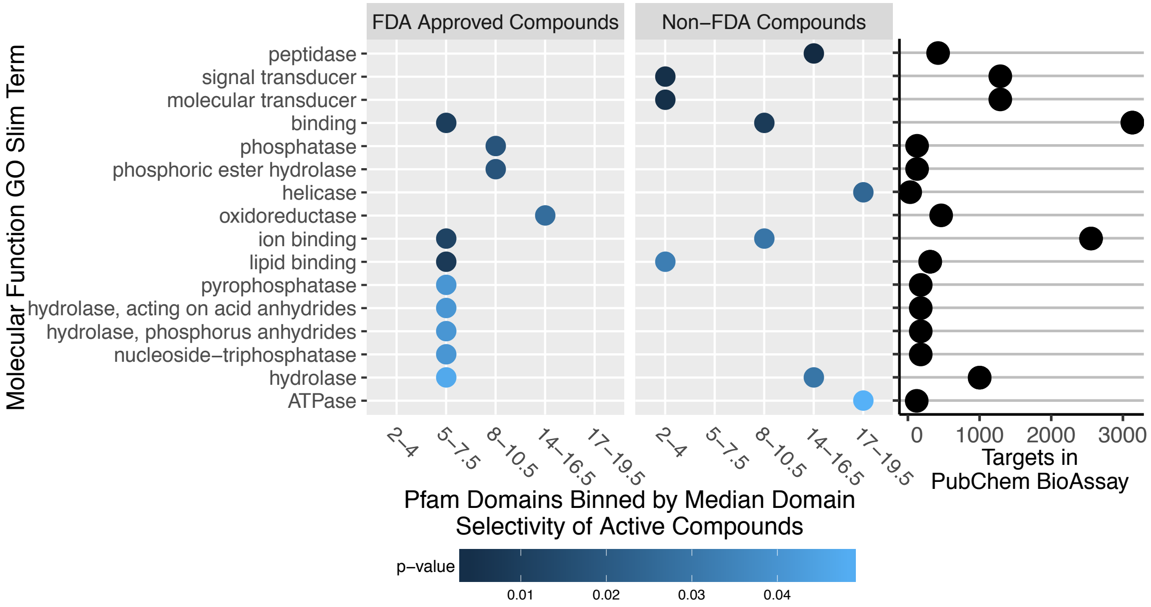


Figure 5.5: **Molecular Function Gene Ontology (MF GO) slim Term Enrichment vs Domain Selectivity.** Pfam domains are binned by the median domain selectivity of active compounds against targets with these domains, as in Table 5.3. The domains in each bin were computed separately based on FDA approved and non-FDA compounds, shown here side by side. For each bin of domain selectivity, the enrichment of MF GO slim terms against the background of all bins is shown. Enriched terms are sorted increasingly by the lowest p-value obtained, with all terms shown here having a p-value < 0.05. The right column dot plot shows the number of protein targets in PubChem BioAssay annotated with each MF GO slim term.

$$P(x) = e^{-(x/x_0)^c} \tag{5.1}$$

Target Selectivity and Compound Complexity

We investigated the distribution of target selectivities across compounds of different molecular sizes, and found that the overall distribution is similar, however very large compounds tend to have fewer active targets, and FDA approved drugs are slightly smaller on average than non-FDA compounds. The definition of molecular size used here is the quantity of non-hydrogen atoms. The very largest FDA approved drugs tend to be natural products, which have several distinct patterns of target selectivity. For example, large antibiotics which evolved to inhibit prokaryotic ribosomal RNA structures tend to be extremely selective or inactive against protein targets, whereas many natural antimitotic and antiparasitic molecules are highly promiscuous. We present additional results from the molecular size target selectivity analysis in Supporting Information.

5.3.3 Promiscuous Binders and Hit Ratio Statistical Model

Hit Ratio Model

Cross-reactive or "promiscuous" compounds are regarded as problematic in drug discovery efforts, as they show activity in a large fraction of HTS experiments, but fail to exhibit selective activity against the desired biological target(s) [36, 2, 15, 8]. Here we model the probability of a compound being promiscuous by estimating the hit ratio, θ with Bayes' rule, based on it's individual screening data. Hit ratio is the expected fraction of active targets that would be found if a compound were screened against the full target space represented in PubChem BioAssay. We model hit ratio with a binomial distribution, using a beta distribution conjugate prior in the manner developed by Dančík, V *et al.* (see "Hit Ratio Bayesian Model and Mixture Distribution" Methods section) [8]. This method enables filtering, and comparative ranking of compound promiscuity unbiased by individual assay participation.

By taking an equal number of random samples from the hit ratio posterior distributions for a set of compounds, we generate an equally weighted convex combination of hit ratio probabilities. This represents the probability of any individual compound from a set having a specific hit ratio, allowing us to compare the evidence for different hit ratios across different compound sets. Here we investigate the promiscuity of FDA approved drugs vs non-FDA compounds, and also investigate the ability of two common methods of identifying promiscuous compounds, pan-assay interference compound (PAINS) functional groups and promiscuous aggregator assays to distinguish between compounds that show selective vs promiscuous behavior in large bioactivity data [36, 2].

FDA Approved vs Non-FDA Compounds

In Fig 5.6A we plot the hit ratio probability distributions for FDA approved and non-FDA compounds, computed as described above. The non-FDA compounds have a high probability density at low hit ratios (left side of plot), whereas the FDA approved drugs have much greater density at high hit ratios (middle and right of plot), consistent with the greater number of active targets described in the above "Target Selectivity Distribution" section. Quantitatively, for individual FDA approved drugs, there is an 85% probability of having a hit ratio below 17.8% ($P(\theta_{approved} < 0.178) = 0.85$), while there is an 85% probability of a non-FDA drug having a hit ratio below a much lower threshold of 3.27% ($P(\theta_{other} < 0.0327) = 0.85$). We also performed a two-sample Kolmogorov-Smirnov test, which measured a distance of $D = 0.465$ between the two probability distributions, indicating that the FDA approved and non-FDA compounds have nearly half of their probability density at different hit ratios. This test metric has a range between 0 and 1, indicating the maximum distance between the cumulative sums of the two probability distributions. Both the FDA approved drugs, and non-FDA compounds show a multimodal distribution dominated by highly selective compounds (left side of plot, approximately $\theta \leq 0.05$), a tail of middle range selective compounds (middle of plot, approximately $0.05 < \theta \leq 0.55$), and a portion of promiscuous binders (right side of plot, approximately $0.55 < \theta$). The promiscuous binder tail among FDA approved compounds is dominated by drugs with well known promiscuous activity such as dasatinib (active against 145 out of 204 screened targets in the PubChem BioAssay data), sunitinib (active against 272 out of 313 screened targets), and morphine (active against 15 out of 16 screened targets) [44]. As cancers tend to exhibit robustness against inhibition of individual kinases, compounds which exhibit broad polypharmacology across the kinases are widely utilized in clinical oncology [28].

Promiscuous Aggregators

Promiscuous aggregators are small molecules that pose a significant challenge to high throughput screening, as they form colloidal aggregates that nonspecifically inhibit enzymes and other protein targets [36, 12]. To assess the ability of experimentally identified aggregators to distinguish between compounds with a high vs low hit ratio, we computed the hit ratio probability distributions separately for promiscuous aggregators and non-aggregators, as shown in Fig 5.6B. To facilitate this, we obtained

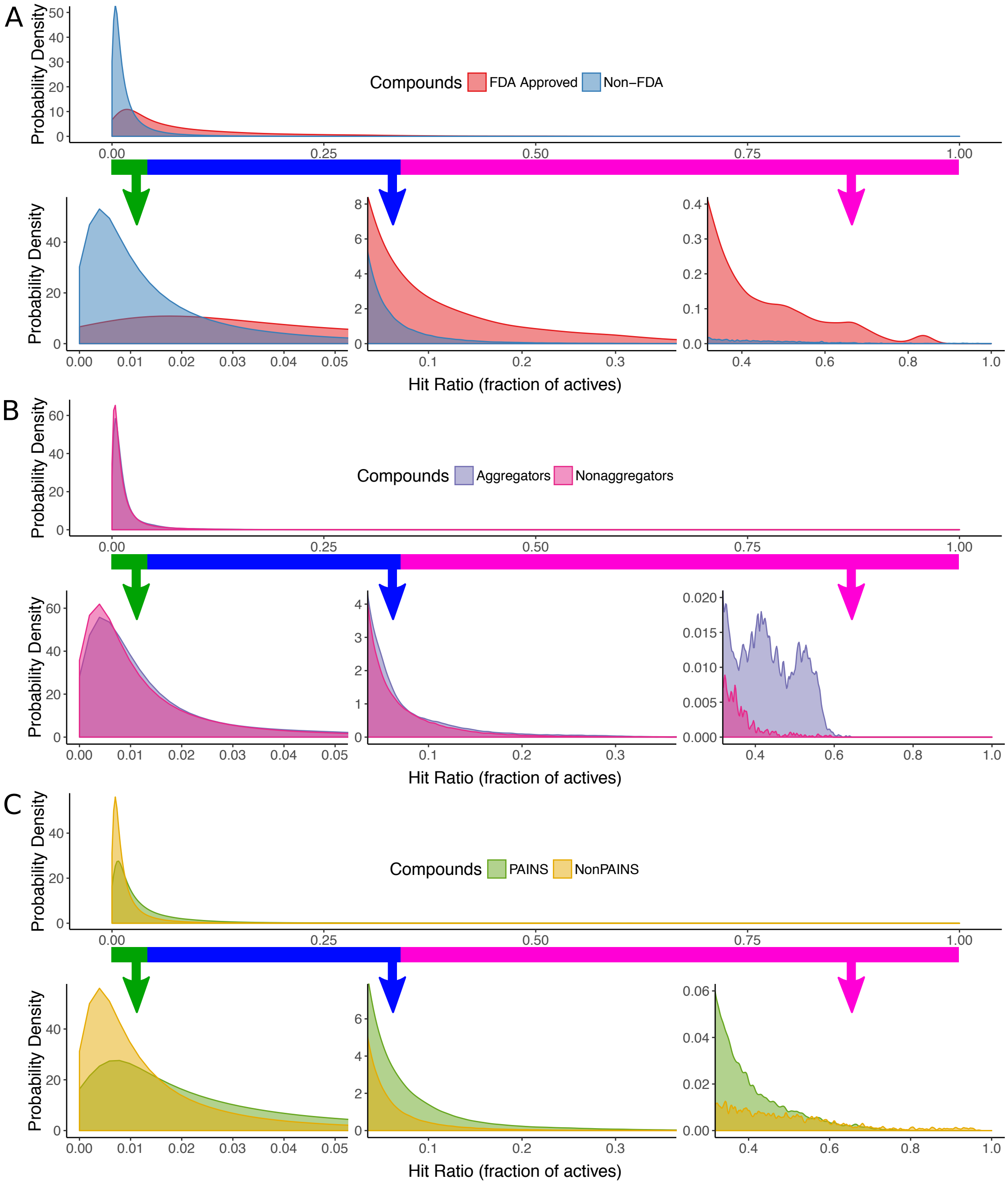


Figure 5.6: **probability density of hit ratios** (θ) Shown here is an equally weighted convex combination of hit ratio probabilities for individual compounds, which represents the probability of any individual compound from a set having a specific hit ratio. Smoothing was applied to reduce sampling noise in low probability regions. **(A)** Hit ratio distributions for FDA approved compounds vs non-FDA. **(B)** Hit ratio distributions for aggregator compounds vs non-aggregators. **(C)** Hit ratio distributions for PAINS vs non-PAINS.

a list of 1185 highly screened active aggregators and 55248 highly screened active nonaggregators previously identified by detergent-dependant inhibition of AmpC β-lactamase as reported by Feng *et al.* (see “Promiscuous Aggregators” in Methods) [12].

For aggregator compounds there is an 85% probability of having a hit ratio below 3.18% ($P(\theta_{approved} < 0.0318) = 0.85$), while there is an 85% probability of a non-aggregator having a hit ratio below a slightly lower threshold of 2.73% ($P(\theta_{other} < 0.0273) = 0.85$), showing that aggregators tend to be more promiscuous across the PubChem BioAssay data, but by a small margin. The maximum distance between the hit ratio probability distributions in cumulative probability space is $D = 0.0596$ as measured by a two-sample Kolmogorov-Smirnov test, demonstrating that a majority of aggregators and nonaggregators have a very similar overall hit ratio distribution compared to the distance of 0.465 between FDA and non-FDA drugs reported in the previous section. However, aggregators show high fidelity in identifying highly promiscuous compounds, as shown in the upper range of hit ratios in Fig 5.6B (right side of plot). This indicates that while most of the aggregators fail to show promiscuous activity across the PubChem BioAssay data, a large fraction of the most promiscuous compounds are identified as aggregators. We further investigate the ability of aggregators to identify promiscuous compounds in the “Sensitivity and Specificity of Aggregators and PAINS” section below.

Pan-assay Interference Compound (PAINS)

Pan-assay interference compounds (PAINS) are small molecules with substructural features that have been found to exhibit promiscuous activity across many high throughput screens, and may interfere with drug discovery efforts designed to identify target selective compounds [2]. We computed the hit ratio probability distribution separately for PAINS vs non-PAINS as shown in Fig 5.6C. We used the RDKit software library to identify 19988 PAINS compounds, and 298166 non-PAINS compounds, among the set of highly screened actives in PubChem BioAssay (see "Pan-Assay Interference Compounds" in Methods).

For individual PAINS compounds there is an 85% probability of having a hit ratio below 6.60% ($P(\theta_{approved} < 0.0660) = 0.85$), while there is an 85% probability of a non-PAINS compound having a hit ratio below a lower threshold of 3.08% ($P(\theta_{other} < 0.0308) = 0.85$), showing that PAINS tend to be more promiscuous than non-PAINS. The maximum distance between the hit ratio probability distributions in cumulative probability space is $D = 0.228$ as measured by a two-sample Kolmogorov-Smirnov test, demonstrating that PAINS have just under one quarter of their probability density at different hit ratios than non-PAINS. However, compared to the aggregators in Fig 5.6B, they show lower fidelity in identifying highly promiscuous compounds represented in the upper range of hit ratios (see right side of both plots). This suggests that promiscuous aggregators and PAINS may have mutually complementary utility for informing the curation of drug discovery libraries, as we investigate further in the next section.

By comparing the probability distributions in Fig 5.6A and 5.6C, PAINS have a hit-ratio distribution similar to, but somewhat less promiscuous than the FDA approved drugs. This raises a concern, as PAINS are most frequently used to eliminate non-viable drug candidates. However, we find that PAINS have a mean target selectivity count of 8.09 (median 4), but a median domain selectivity count only slightly lower, at 6.72 (median 4). Therefore, for PAINS compounds which are active against many targets, a substantially smaller fraction hit targets with common domains, as compared to the FDA approved drugs, as described in the above "Target Selectivity Distribution" section. This highlights a fundamental difference between PAINS and FDA approved drugs. While both tend to have activity against many targets, PAINS tend to be active against targets with unrelated sequences, while FDA approved drugs tend to be active against related targets.

Sensitivity and Specificity of Aggregators and PAINS

The highly screened active compounds can be divided into promiscuous and non-promiscuous categories based on the evidence in PubChem BioAssay, by choosing a promiscuity probability cutoff where $P(\theta \geq 0.25) >$ cutoff. The number of promiscuous compounds at each cutoff is shown in the lower panel of Fig 6.6 in Supporting Information. For a given cutoff fraction, based on the public bioactivity data, our model predicts that approximately this fraction of compounds classified as promiscuous will have a true hit ratio above 0.25. We assessed the sensitivity (true positive rate) and specificity (true negative rate) of both PAINS and aggregators to categorize promiscuous compounds throughout a range of cutoffs from 0.01 to 0.9999. Here, sensitivity is defined as the fraction of compounds classified as promiscuous at a given cutoff that are also identified as PAINS or aggregators respectively, while specificity is the fraction classified as non-promiscuous that were also identified as non-PAINS or non-aggregators respectively.

PAINS showed a maximum sensitivity of 21% at a cutoff of 0.08, and aggregators showed a maximum sensitivity of 38% at a cutoff of 0.9996. Both non-PAINS and non-aggregators had a nearly constant specificity throughout this range, with non-PAINS having a specificity of 94%, and non-aggregators having a higher specificity of 98%.

As shown in the upper panel of Fig 6.6 in Supporting Information, the two have opposite trends where PAINS show decreasing sensitivity at increasing promiscuity cutoffs, while aggregators show increasing sensitivity at higher cutoffs. This is consistent with the probability distributions in Fig 5.6, in that both identify compounds with high hit ratios, but the PAINS compounds are more enriched in the middle range of hit ratios, while the aggregators tend to be highly promiscuous. While aggregators show both higher sensitivity and higher specificity, they identify a much smaller subset of promiscuous compounds that have extremely high hit ratios, and are not able to identify the large number of compounds with middle-range hit ratios that PAINS identifies.

Using a promiscuity probability cutoff of 0.5, our statistical model found 1409 promiscuous binders among the entire highly screened active PubChem BioAssay compound set, as shown in the center of Fig 6.6 in Supporting Information. Of these promiscuous binders, 1157 are not currently included among the set of PAINS or aggregators used here, and 75 are FDA approved drugs. The number of FDA approved drugs reduces to 24 with a higher promiscuity probability threshold of 0.999. As a resource for readers, we include the computed promiscuity probabilities $P(\theta \geq 0.25)$ for all highly screened actives in Supporting Information S2 File, sorted by decreasing probability of promiscuity. This also serves to rank the compounds by target selectivity, with a ranking that is meaningful based on the experimental evidence, despite varying levels of assay participation.

5.3.4 Comparison Between Annotated Drug Targets and Public HTS Data

We systematically compared the bioactivity data in PubChem BioAssay with the annotated targets of FDA approved drugs in DrugBank (version 4.2), in order to assess the level of agreement between the two, and identify the number of potential novel targets for the FDA approved drugs [52]. To enable this, we created a drug-target matrix encoding both bioactivity data and target annotations in a directly comparable manner. The rows represent the highly screened FDA approved drugs, while the columns represent all of the PubChem BioAssay screened and DrugBank annotated targets for these compounds. As many PubChem BioAssay activity results were generated with truncations of endogenous proteins, or using close orthologues to putative *H. sapiens* targets from other species, it was necessary to merge data from very similar targets, as described in the “Clustered Compound-Target Matrix” methods section. This resulted in 1829 distinct protein target columns, of which 1416 have a *H. sapiens* representative UniProt identifier. Each compound-target pair (position) in the matrix was assigned one of six possible values depending on it's DrugBank annotation (annotated vs. unannotated) and it's PubChem BioAssay activity results (untested, active, inactive). The resulting comparison between the BioAssay data and DrugBank annotations is shown in Table 5.6. There is a high level of agreement between the DrugBank target annotations and the PubChem BioAssay data, with 1082 compound-target pairs in agreement, and only 83 compound-target pairs in disagreement, where they are annotated as active in DrugBank but were found inactive in PubChem BioAssay. While the matrix is very sparse, with the majority of compound-target pairs both unscreened and unannotated, the PubChem BioAssay data substantially increases the density of the compound-target matrix, with 7817 active compound-target pairs not present in the DrugBank annotation, representing a new space of potential targets for these drugs. There are 867 protein target clusters (751 *H. sapiens*) that are annotated as active within DrugBank, however an additional 804 protein target clusters (576 *H. sapiens*) show activity in PubChem BioAssay but have no existing DrugBank annotation. Some of these active but currently unannotated targets may represent new target space that can be used to repurpose existing drugs for novel therapeutic purposes, or to explain currently unknown or unannotated targets in existing therapies. We provide a full list of these potentially novel drug-target pairs in S3 File of Supporting Information.

While the false positive rate of these PubChem BioAssay activity outcomes is not precisely known, our estimate above using replicated assay pairs suggests that

	Unannotated in DrugBank	Annotated Target in DrugBank
Untested in PubChem	1431855 (1111148 *H. sapiens*)	2097 (1900 *H. sapiens*)
Inactive in PubChem	153783 (115181 *H. sapiens*)	83 (83 *H. sapiens*)
Active in PubChem	7817 (6848 *H. sapiens*)	1082 (1008 *H. sapiens*)

Table 5.6: **Comparison of PubChem BioAssay activity data to DrugBank target annotations.** All compound-target pairs for FDA Approved drugs are grouped into one of six possible categories. Depending on the HTS results in PubChem BioAssay, a compound-target pair is annotated as either untested, inactive, or active (rows in this table). Additionally, the compound-target pair is either annotated or unannotated as a known active target in DrugBank (columns in this table). Counts outside of parenthesis represent results against all protein targets, whereas counts to the right in parenthesis represent results against the subset in which the representative UniProt indentifer for each target cluster is from the *H. sapiens* proteome.

the number of false positives is less than, and of the same order of magnitude as the number of true positives. Consequently, we predict that at least half of these novel drug-target activity results are experimentally repeatable. Additionally, as we demonstrate in the next section, many of these new currently unannotated active values fall into dense biclusters, where the same compound has been found active against a large number of closely related protein targets across many assays. As these biclusters are highly enriched for a large number of active scores, these are unlikely to be a result of random error. We provide a full list of these high confidence biclusters in S4 File and S5 File of Supporting Information.

5.3.5 Drug-Target (DT) Biclustering Analysis

Biclustering Overview

To investigate the possibility of shared patterns of activity between sets of FDA approved drugs and their protein targets, we created a drug-target binary activity matrix based on the drug-target matrix in the above section. Each drug-target combination was assigned a value of 1 if active in PubChem BioAssay, or annotated as a known target in DrugBank. Untested or inactive values are assigned a value of 0. The resulting bioactivity matrix has a total active in PubChem BioAssay and/or annotated as active in DrugBank score density of 0.69%. We then clustered this matrix using the BicBin sparse biclustering algorithm (see methods) [46]. This type of clustering algorithm clusters rows and columns simultaneously allowing us to identify both sets of compounds and targets sharing similar activity profiles within each dimension. BicBin was chosen among several biclustering algorithms as it finds sparse biclusters with flexible options, scales to large matricies, and finds top-scoring clusters first. We identified the 16 highest scoring biclusters which contained at least two compounds and at least two targets as shown in Table 5.7. These biclusters had an activity density substantially higher than the entire matrix, ranging from 31.46% to 92.19%. These biclusters contain 406 drugs, of which 136 appear in multiple biclusters with a maximum of 6 biclusters per compound, and 346 unique representative protein targets, of which 107 appear in multiple biclusters with a maximum of 4 biclusters per protein. Fig 6.4 in Supporting Information shows the entire matrix represented as a bipartite graph with compounds colored by their highest scoring bicluster (white if unclustered), and protein targets in black. We found that the very sparsely connected graph clusters into very densely connected biclusters where a sizable set of distinct drugs has been found to be active against a sizable shared set of distinct targets. To functionally annotate each bicluster, we identified the most common Pfam domains present in their protein targets [41]. As shown in Table 5.7, in some biclusters most or all of the protein targets share a common domain that is the likely the target of these compounds, *e.g.* 22 out of 27 targets in bicluster 1 share the rhodopsin-like receptor domain (PF00001: 7 transmembrane receptor). Fig 5.7 visualizes the compound-target activities in bicluster 1 as a heatmap. In other biclusters, only a small fraction share a common domain such as bicluster 15 where only 6 out of 57 targets share PF00001. These cases warrant deeper investigation as to why they share a common activity pattern, but with a more heterogeneous domain composition. This questions is investigated in the next section. As mentioned in the previous section, we provide a full list of these biclusters in S4 File and S5 File of Supporting Information.

#	Compounds	Targets	Top Pfam Domain	W/ Domain	Score
1	62	27	PF00001 7 transmembrane receptor (rhodopsin family)	22	13.35
2	130	5	PF00067 Cytochrome P450	5	11.61
3	119	9	PF00104 Ligand-binding domain of nuclear hormone receptor	6	9.52
4	53	18	PF00194 Eukaryotic-type carbonic anhydrase	9	7.28
5	4	97	PF00069 Protein kinase domain	89	7.28
6	10	24	PF00001 7 transmembrane receptor (rhodopsin family)	13	5.97
7	48	23	PF00001 7 transmembrane receptor (rhodopsin family)	15	5.61
8	63	28	PF00001 7 transmembrane receptor (rhodopsin family)	7	4.79
9	2	67	PF00069 Protein kinase domain	49	4.46
10	2	78	PF00001 7 transmembrane receptor (rhodopsin family)	10	4.05
11	8	8	PF00484 Carbonic anhydrase	6	4.76
12	6	13	PF00520 Ion transport protein	4	3.91
13	57	4	PF00001 7 transmembrane receptor (rhodopsin family)	2	4.36
14	18	5	PF00001 7 transmembrane receptor (rhodopsin family)	5	4.78
15	2	57	PF00001 7 transmembrane receptor (rhodopsin family)	6	3.89
16	12	16	PF00817 impB/mucB/samB family	3	3.84

Table 5.7: **Top Pfam domains in each bicluster.** Shown are the top 16 highest scoring drug-target biclusters with more than one compound and more than one target. The number of drugs (cids) and targets is shown in columns 2 and 3, respectively. The 4th and 5th columns give the name of the most abundant domain and its frequency, respectively. The last (6th) column shows the BicBin score, representing the density and size of the bicluster. The BicBin score is the negative exponent of the Chernoff Bound. It is inversely proportional to the probability of each bicluster occurring by random chance, as described in Methods.

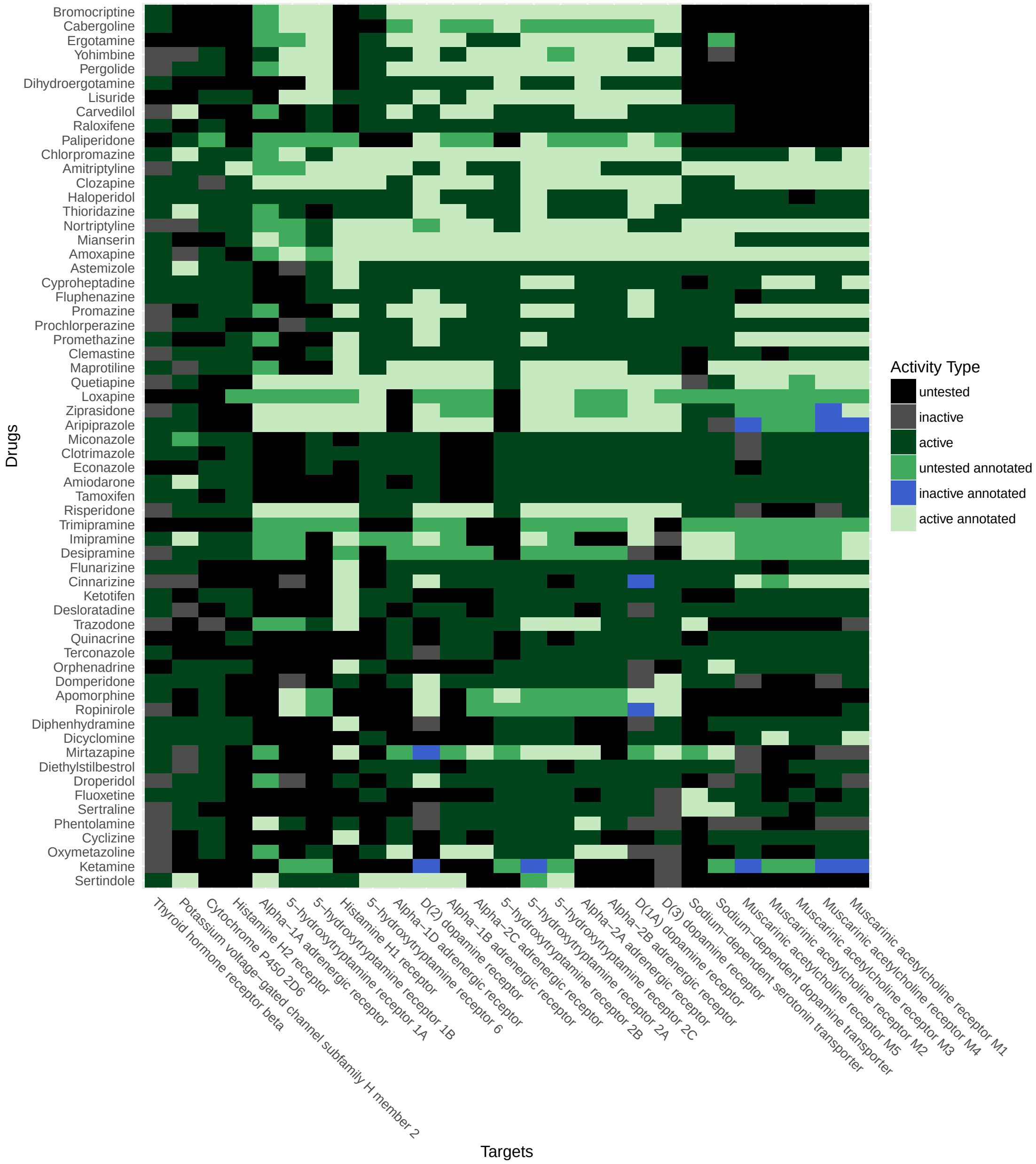


Figure 5.7: **Bioactivity of drug-target bicluster 1.** The vertical axis lists the drugs in this bicluster by common name, and the horizontal axis represents the UniProt names for the representative targets of each sequence-similar target cluster. The compound-target pairs are colored according to one of six colors: untested in PubChem BioAssay (black), inactive in PubChem BioAssay (grey), active in PubChem BioAssay (dark green), untested but annotated as active in DrugBank (green), inactive in PubChem BioAssay but annotated as active in DrugBank (blue), and active and also annotated as active in DrugBank (light green). Rows and columns are sorted by bioactivity profile similarity.

Drug-Target (DT) Bicluster GO Slim Analysis

To further categorize each bicluster by functional processes, we performed an enrichment analysis of the Molecular Function GO Slim terms associated with the representative protein targets within each bicluster. Most biclusters exhibit a distinct pattern of enriched GO terms, distinguishing them from other biclusters. For example, bicluster 5 consists of four kinase inhibitor drugs with known broad kinase-activity (Dasatinib, Sorafenib, Erlotinib, and Gefitinib), and a highly enriched kinase GO term (PF00069, p-value $9.11 * 10^{-64}$) present in the annotation of 89 out of a total of 97 targets in this bicluster. Additionally if the entire drug-target network is colored by the GO terms present in each target, a distinct regional pattern emerges, where targets sharing active compounds also tend to share common GO terms as shown in Fig 6.5 of Supporting Information. Overall the GO Slim annotations provide a more informative functional summary of each bicluster than the Pfam annotations. This is often the case because a greater portion of the targets in each bicluster tends to share the most enriched GO term, but not necessarily a specific Pfam domain.

As the top scoring biclusters listed here include approximately half of the highly screened active FDA approved compounds, they constitute specific examples which partially explain the higher number of active targets among FDA approved drugs compared to non-FDA compounds (consider Fig 5.4, as well as the greater probability density at higher hit ratios in Fig 5.6A). In summary, a substantial fraction of the FDA approved compounds show broad activity across a large set of related targets

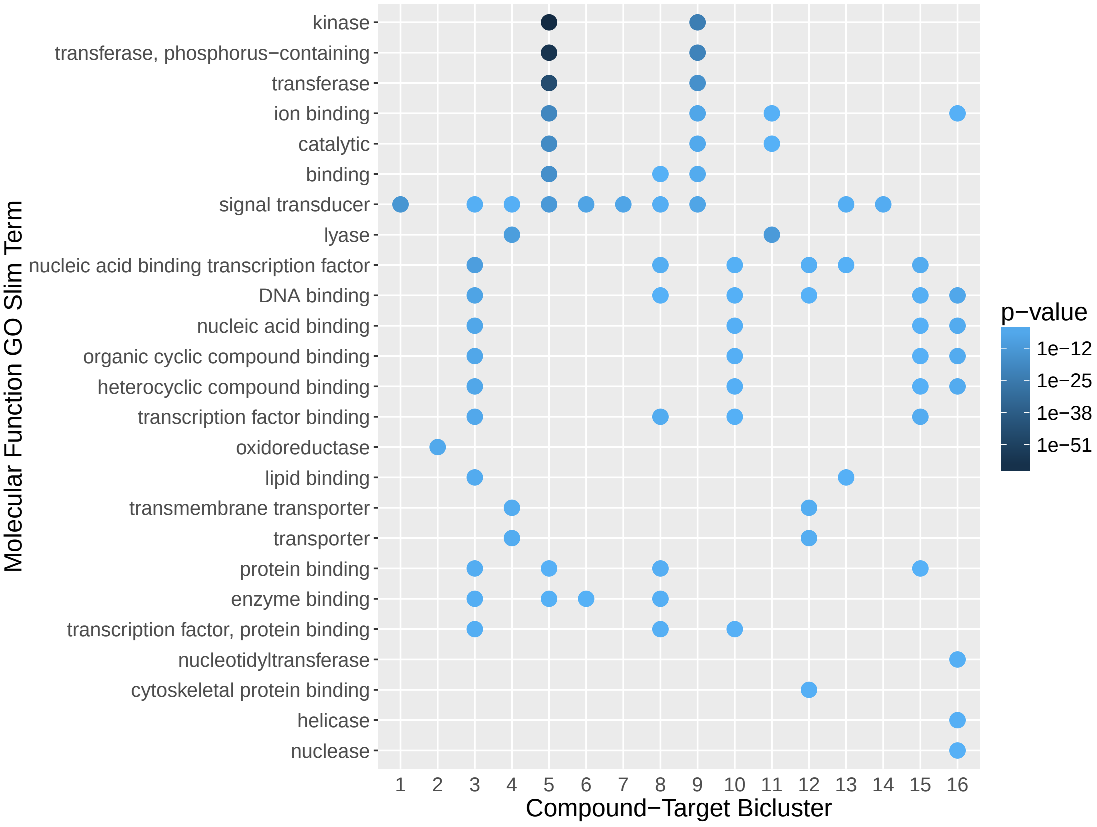


Figure 5.8: **Molecular Function Gene Ontology Slim (MF GO Slim) term enrichment for each drug-target bicluster.** Enrichment measured by hypergeometric test. Terms with $p \leq 0.05$ are shown and sorted increasingly.

in the same bicluster, which are enriched for common Pfam domains and/or MF GO slim terms.

Compound Structure vs Bioactivity Bicluster Analysis

In order to compare the compound structure vs bioactivity patterns among these biclusters, we clustered the FDA approved drugs by structural similarity using atom pair (AP) descriptors and the Tanimoto coefficient as similarity metric [4]. Fig 5.9 shows the compounds from the 11 largest bioactivity biclusters, positioned according to structural similarity, and colored according to their lowest numbered (densest and/or largest) bioactivity bicluster. The structural distances were used to project each compound into two dimensional plane with multi-dimensional scaling (MDS) where the points (compounds) are spaced proportionally to the chemical structure difference between the compounds, with more similar compounds closer together. Visually, two distinct patterns can be identified where structurally similar compounds (in close proximity) also cluster together with similar bioactivity (*e.g.* bicluster 1 shown in light blue, a cluster of primarily aromatic compounds active against G-protein-coupled receptor targets). However, the opposite can also be observed, where compounds with very similar bioactivity have diverse structures (*e.g.* bicluster 3 shown in light green, a cluster with many nuclear receptor targets).

To estimate the extent to which the structure-function principle (*i.e.* that similar structures have similar bioactivities) applies to this data, we clustered the 406 compounds represented in the biclusters into discrete clusters using complete linkage hierarchical clustering with subsequent tree cutting with $k = 11$. The latter value matches the number of biclusters remaining when the compounds are each assigned to a unique bicluster. To quantify the similarity among the structural clusters and the bioactivity clusters from of the above biclustering section, we compared the numbers of identical and unique compound pairs appearing in the two clustering results using the Jaccard index. The result indicated that 15.21% of compound pairs were joined into clusters by both methods. If the structural clustering is replaced with a random grouping into one of 11 clusters weighted by the cluster size distribution in the structure clustering, we see a mean Jaccard index of only 11.10% ($sd = 0.39\%$

and permutation p-value 0.0001) across 10,000 random clusterings. This quantifies what can be seen visually in Fig 5.9, that overall structural similarity correlates with bioactivity similarity, but with a sizable number of exceptions.

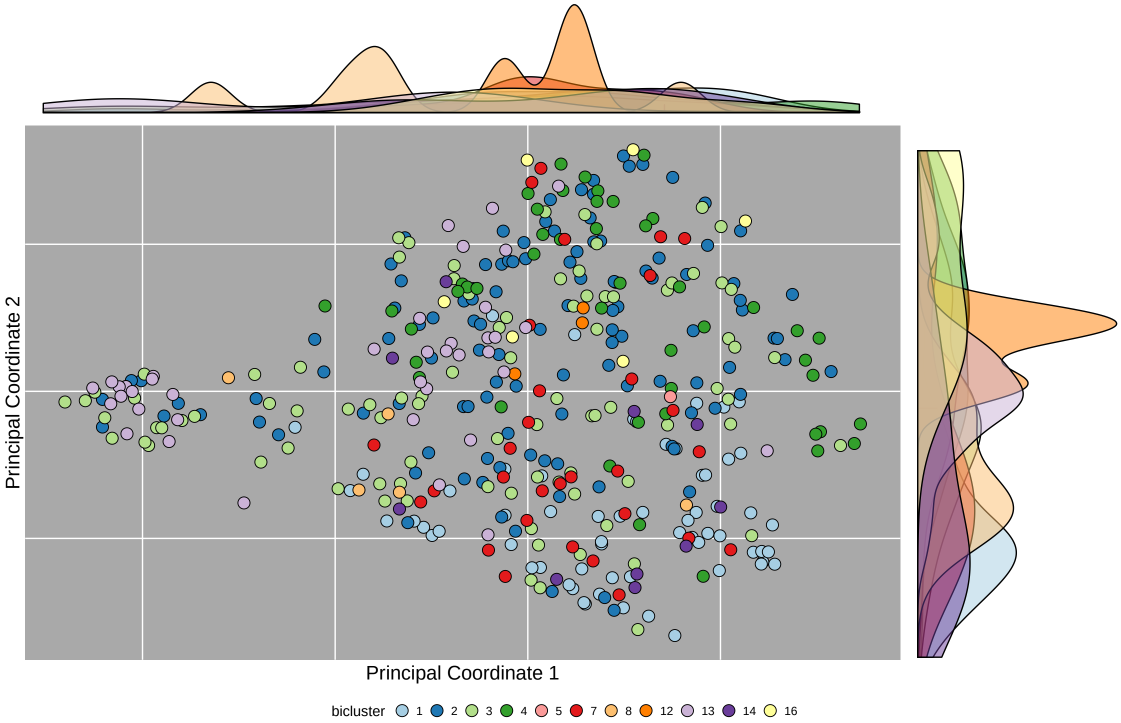


Figure 5.9: **Bicluster (color) vs compound structure (position).** Multidimensional scaling (MDS) was used to embed small molecules into a two dimensional space (x- and y-axis). Each point represents an FDA approved drug. A density map colored by each bicluster is shown for the MDS principal coordinate 1 (on top) and principal coordinate 2 (on right). The distance between the points is proportional to the chemical similarity between the two compounds. Bioactivity-based biclustering results are also indicated by colors, with each compound assigned to it's lowest numbered (densest and/or largest) bicluster. Only the 11 biclusters with the largest number of compounds are shown, to allow for a visually distinct color palette.

5.3.6 Target-Protein (TP) Network

In order to extend the drug-target biclustering analysis shown above to the full set of PubChem BioAssay bioactivity data, we created a Target-Protein (TP) network where proteins are connected if they are targeted by over 50% of the same non-promiscuous compounds (as described in Methods). This was inspired by the TP network previously published by Yildirim *et al.*, while adding a bioactivity similarity threshold and excluding promiscuous compounds, in order to enable the incorporation of large primary screening data while limiting spurious edges [54]. This graph approximates the structure of a full compound-target binary activity matrix (or bigraph), in a computationally efficient manner by excluding the small molecule nodes.

The final graph is shown in Fig 5.10. It has 2407 nodes (target proteins with at least one edge) and 11317 edges. There are 176 connected components with the majority of nodes in the largest. The average degree is 9.40 with a graph density of 0.004. In Fig 5.10, protein targets are colored according to the 11 most abundant Molecular Function Gene Ontology Slim (MF GO Slim) terms among the PubChem BioAssay protein targets, with nodes lacking any of these 11 terms labeled in black (other). If a target is annotated with more than one of these 11 terms, the most specific term was chosen as the representative color for that node. Fig 5.10 demonstrates a distinct grouping of protein targets with shared bioactivity by MF GO terms. As the structure of the network was not informed with MF GO annotations, this indicates that the overall pattern demonstrated above for the drug-target biclustering analysis also extends to the full set of compounds and targets in the PubChem BioAssay data, in that targets sharing a common MF GO slim annotation tend to have a distinct but shared set of active small molecules.

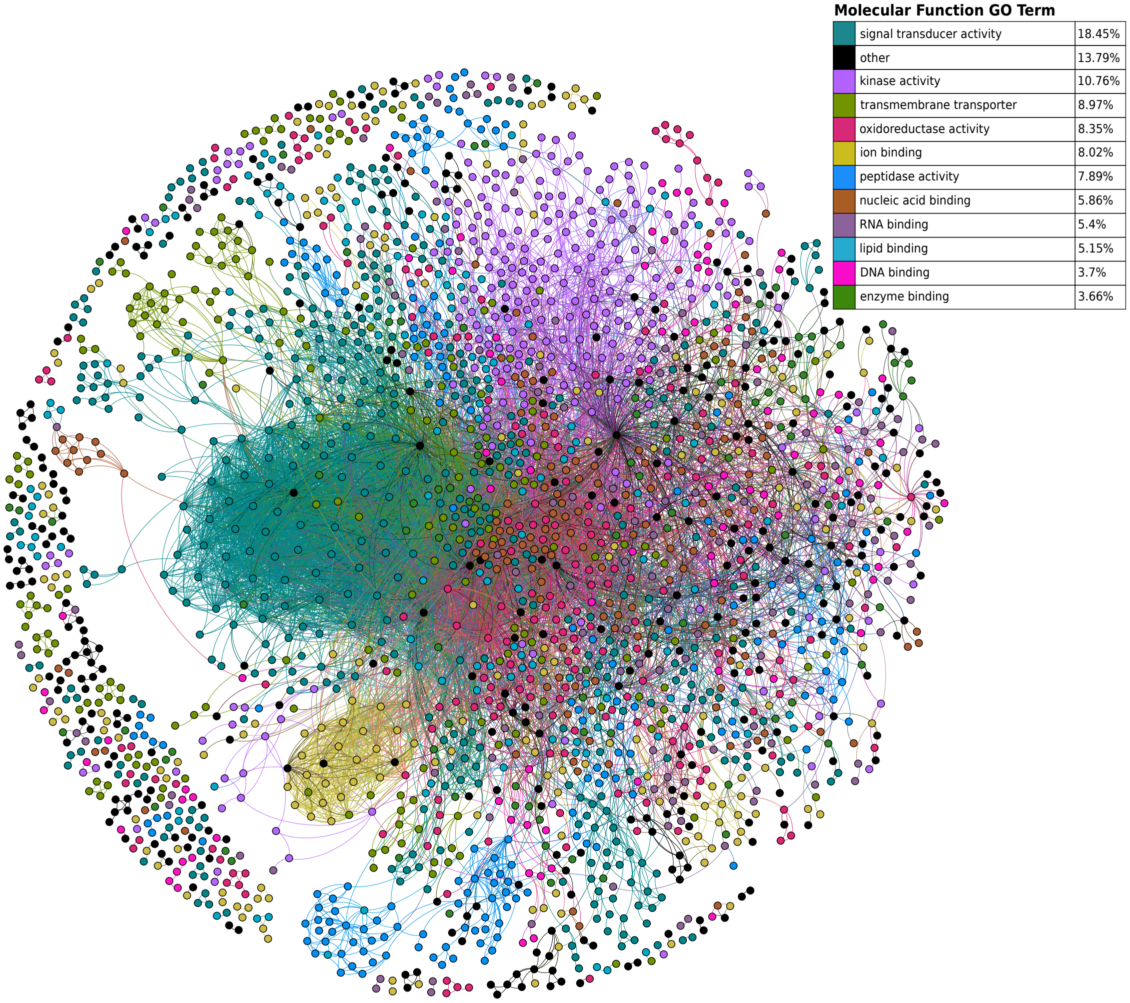


Figure 5.10: **Target-protein network and Molecular Function Gene Ontology Slim (MF GO Slim).** Each node represents a protein target, and edges connect any two protein targets with greater than 50% bioactivity similarity across non-promiscuous binding compounds. Targets are colored according to MF GO Slim terms, with unannotated targets colored black. Shown are 2407 nodes (target proteins with at least one edge) and 11317 edges representing shared bioactivity among the mutually screened subset of the 1.2M compounds tested in the bioassays we analyze in this study.

5.4 Methods

Most analysis steps were performed with the open source software R, bioassayR and ChemmineR. The latter two are Bioconductor packages developed by the authors. Several of the bioactivity methods are described in Backman *et al* [1]. The full source code of the analysis presented in this paper is freely available online at `http://github.com/girke-lab/targetSelectivity`

5.4.1 Bioactivity Database

We used the R package bioassayR to build a database which contains all small molecule bioactivity screens from PubChem BioAssay which include at least one real activity score (active or inactive) and have a single protein target specified. Both raw numeric scores, and discrete active/inactive categories were parsed, and stored in the database, however direct cross-comparison between the numeric scores is limited by varying assay designs and scoring methods.

5.4.2 Clustering Protein Targets by Sequence

We used the kClust tool to cluster the non-redundant set of both the protein targets in PubChem BioAssay as well as the protein targets interacting according to DrugBank with FDA approved drugs [17, 52]. Stringent threshold settings were chosen to merge very close orthologues, paralogs, and engineered proteins (*e.g.* truncations performed for screening purposes). The parameters used were $s = 2.93$, E-value $\leq 10^{-4}$, $c = 0.8$. For each resulting cluster, a single representative protein was chosen for annotation purposes. These representatives were chosen with the following order of precedence: an annotated *H. sapiens* drug target (from DrugBank), any *H. sapiens* target with a known UniProt identifier, a non *H. sapiens* target with a known UniProt identifer, and lastly a non *H. sapiens* target with only a GenBank GI number and no known UniProt identifier.

5.4.3 Clustering of Compound-Target Matrix

We used bioassayR to generate a compound-target binary sparse matrix summarizing a substantial fraction of the protein target bioactivity data in PubChem BioAssay. Only compounds screened against at least 10 protein targets (distinct GenBank GIs) were included, in order to avoid biasing the selectivity analysis by compounds with too limited data. In order to reduce the sparseness and duplication in these data, assays sharing identical protein targets, or targets falling into the same sequence cluster (see above) were merged into a common column. The merging was performed in a way where active scores take precedence over inactive scores. Each column was annotated by a representative protein for that target cluster, as described above.

5.4.4 Protein Annotations and GO Enrichment

Pfam-A (version 29.0) domains were mapped to target proteins with HMMER3 (version 3.1b2) [41, 10]. An E-value ≤ 0.01 was used as domain reporting threshold. The target proteins included all PubChem BioAssay targets, DrugBank annotated targets, and the *H. sapiens* reference proteome (proteome ID UP000005640) provided by UniProt [6]. Gene Ontology annotations for protein targets were obtained from UniProt, while Gene Ontology annotations for Pfam domains were obtained from InterPro [37]. The subset of GO slim terms (Generic version) was obtained from the GO Consortium. Hypergeometric GO term enrichment tests were performed using the R language GOstats and GSEABase packages [11].

5.4.5 BicBin Biclustering Of Drug-Target Matrix

The BicBin algorithm was used to identify bioclusters iteratively using the parameters $\alpha = 0.6$, $\beta = 0.6$, representing no bias between adding compounds or targets [46]. These thresholds were chosen to find the largest possible biclusters, without merging biclusters that share little or no overlapping activity. The BicBin biclustering algorithm used here finds dense biclusters of compound-target activity by scoring them with the multiplicative version of the Chernoff Bound applied to the Binomial distribution, which estimates the upper limit of the probability of these clusters occurring by random chance [46]. The bicluster scores shown in Table 5.7 represent the negative exponent of the Chernoff bound, and therefore higher scores correspond to lower probabilities, and therefore larger and denser biclusters. Biclusters were found and scored iteratively, by first zeroing out the previous biclusters.

5.4.6 Hit Ratio Bayesian Model and Mixture Distribution

We model hit ratio θ of each compound with a binomial distribution, using a beta distribution conjugate prior in the manner developed by Dančík, V *et al* [8]. For a given number of active targets n, out of N screened targets, we assume that n has a binomial distribution, as in Eq 5.3. We then apply Bayes' rule (Eq 5.2) to compute the posterior $P(\theta|n)$ with a beta distribution conjugate prior as in Eq 5.4. The values of α and β are computed from the mean μ and standard deviation σ of hit ratios among all active compounds screened against at least 20 distinct targets, using the bioassayR function crossReactivityPrior and Eqs 5.5 and 5.6. We found a prior hit ratio mean of 0.0186, with a standard deviation of 0.0349.

$$P(\theta|n) = \frac{P(\theta)P(n|\theta)}{P(n)} \tag{5.2}$$

$$P(n|\theta) = \binom{N}{n}\theta^{n}(1-\theta)^{N-n} \tag{5.3}$$

$$P(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{5.4}$$

$$\alpha = \mu^2\left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) \tag{5.5}$$

$$\beta = \alpha\left(\frac{1}{\mu} - 1\right) \tag{5.6}$$

We then compute the probability of a compound being a promiscuous binder $P(\theta \geq 0.25)$ with the bioassayR function *crossReactivityProbability*, and can obtain random samples from the posterior distribution of each compound with the R language function *rbeta*. To obtain an equally weighted convex combination of hit ratios for a compound set, we took an equal number of samples with *rbeta* for each compound, and then took one million unbiased samples without replacement from these. To plot this distribution we used the *geom_density* function of the ggplot2 software library with the option $adjust = 3$ to smooth sampling noise in the tails.

The Bayesian hit ratio model is based on an underlying assumption that the available activity data for a given compound represents activity against a random sample with replacement of the screenable protein target space. While this is a reasonable approximation for compounds screened against a large number of diverse targets, in many cases compounds screened against a small number of targets are likely to have substantial bias in their target set. Therefore, there is a strong possibility that a compound with only one or two active targets is highly selective, or is an inactive compound with a false positive, resulting in overfitting that would make the computed hit ratio an overestimate. As reported by Jasial *et al.*, these compounds with a small number of active targets are unlikely to exhibit undiscovered promiscuity or activity against many additional targets as they are screened in an increasing number of assays [24]. For this reason, here we model hit ratio primarily to identify compounds with a large number of active targets (polypharmacological and promiscuous compounds), while looking at the absolute number of active targets when investigating highly selective compounds, as shown in the "Target Selectivity Distribution" section. Additionally, by using only highly screened compounds, we avoid both many cases of overfitting, and avoid plotting probability distributions for compounds with highly uninformative data, that would result in a non-localized probability density function.

5.4.7 Promiscuous Aggregators

We obtained a list of known promiscuous aggregator and nonaggregator small molecule PubChem compound identifiers (cids) by referencing PubChem BioAssay assays #584 and #585 as described by Feng *et al.*[12] These assays together identify detergent-dependant inhibitors of AmpC β-lactamase. We obtained the list of promiscuous aggregators by identifying compounds marked as active in the assay without detergent (#585), but inactive in the assay with detergent (#584). The list of nonaggregators includes both inhibitors active in both assays, and noninihibitors inactive in the assay without detergent. We excluded from consideration all compounds which obtained an inconclusive result in either assay, or were not highly screened, having been tested in PubChem BioAssay against less than 10 distinct targets. We also excluded compounds without activity against at least one protein target in PubChem BioAssay. This resulted in a list of 1185 promiscuous aggregators, and 55248 nonaggregators.

5.4.8 Pan-Assay Interference Compounds (PAINS)

We used the RDKit software library (version 2016.03.1) SMARTS based PAINS filters to identify compounds classified by the PAINS filters A, B, or C. These SMARTS filters are based on the SMARTS conversion published by Saubern *et al.* based on the

SLN format filters originally published by Baell *et al.* [42, 2] This identified 19988 PAINS compounds, and 298166 non-PAINS compounds, among the set of highly screened actives in PubChem BioAssay. 68 of the compounds we identified as PAINS are also FDA approved drugs. An additional 7814 compounds had structures we could not parse with RDKit and were excluded.

5.4.9 Target-Protein (TP) Network and Network Visualizations

Targets were connected by bioactivity profile similarity using the *trinarySimilarity* function of bioassayR, with default options. This computes Tanimoto similarity coefficients between bioactivity profiles, by considering only commonly tested compounds. The Tanimoto, as computed here, is the size of the intersection divided by the size of the union of active compounds between the two targets. If the pair of targets did not share at least 12 mutually screened compounds, or at least 3 actives, we categorized this pair as having insufficient evidence, and assigned a similarity value of 0. The similarity matrix was converted to a binary connection matrix based on a similarity value of at least 0.50, and then converted to a graph object with the R package igraph [7]. All network visualizations were generated with Gephi using the ForceAtlas2 layout algorithm [3, 22]. Because the layout engine itself was not provided with any annotation information (color), the color based groupings are solely based on the level of connectivity.

We did not exclude infrequently screened compounds as in the other sections in this study, as this analysis was able to make meaningful use of those compounds. We found the overall structure of the graph is roughly the same at different similarity thresholds, however we chose this higher cutoff to reduce the number of edges in the visualization. This high evidence threshold also avoids spurious edges resulting from false positive activity outcomes. Despite such a high cutoff, the majority of the graph is highly connected, showing that a large number of target pairs share very similar activity profiles across a large number of small molecules. Compounds were excluded if their probability of promiscuous binding was greater than 50% ($P(\theta \geq 0.25) > 0.5$). This resulted in the exclusion of 29179 compounds. As this analysis was not limited to highly screened compounds, this is a much higher number than the quantity of highly screened promiscuous compounds reported above. Out of all protein targets in PubChem BioAssay, only 2249 had at least one edge. Removing this small fraction of promiscuous binding compounds (2.52% of total compounds) substantially reduced the number of edges in the network. The number of node pairs (edges) with a computable similarity (enough shared actives or mutually screened targets) above 50% dropped from 283353 to 194444 and the number of highly similar node pairs we connected with edges dropped from 84298 to 9854.

5.5 Conclusion

By systematically analyzing a large volume of public bioactivity data, we highlight several new patterns of bioactivity that may prove useful for informing drug discovery efforts. We also provide additional context to the previously reported finding that FDA approved drugs are, on average, active against a greater number of targets than non-FDA approved active molecules identified by HTS methods[20, 24]. We show that this greater number of targets is not due to biased assay participation, both by using a statistical model which considers the evidence for each compound individually, and by looking at the mean and median assay participation. While still noticeable, the difference in the number of active targets between FDA approved and non-FDA compounds decreases substantially if proteins sharing very similar amino acid sequences, or common domains are not counted separately.

As demonstrated by the high number of previously unannotated active targets (Table 5.6), and the high density of drug-target activity biclusters (Table 5.7 and Fig 5.8), several sets of FDA approved drugs exhibit activity across a shared set of related targets. Previous literature suggests several plausible explanations for how these drugs may have similar or identical bioactivity profiles, while inducing distinct therapeutic phenotypes *in-vivo*. For example, it has been demonstrated that several common drug target receptor families exhibit biased signaling, where a given receptor can activate a large number of downstream processes, in different ratios unique to a given ligand structure, tissue, and organism state [47, 26]. Additionally, bioavailability and biological compartmentalization can limit the *in vivo* access of a small molecule drug to only a small fraction of the targets it many show activity against *in vitro* [33]. Complex network effects and biological feedback can also cause a drug interacting with multiple targets to exhibit functional selectivity. For example, Lehar *et al.* (2009) published an analysis of synergistic drug combinations, showing that combinations of multiple drugs acting against different targets in the same pathways tend to induce a phenotype at lower doses, with lower incidence of off-target effects [32]. Lastly, binary active/inactive HTS data may fail to resolve different receptor binding kinetics that would cause a drug to exhibit target selectivity in the context of a specific dosage level. For example, drugs are often classified and evaluated according to therapeutic index (TI), or the ratio between the dose that results in toxicity to the dose that produces a desired efficacy [39]. For low TI drugs where the desired effect and toxicity are mediated with different receptors that have only slightly different binding affinities, binary active/inactive data could be expected to report activity for both the therapeutic and toxic targets.

This cross-reactivity we observe in FDA approved drugs raises the question and possibility of exploiting this pattern to identify viable drug candidates in noisy and error prone HTS data. With false positives occurring at the same order of magnitude as true positive bioactivity outcomes, it is likely that a substantial fraction of singular active values are due to experimental error. For some drug discovery efforts against target classes where the FDA approved drugs tend to show cross-reactivity within a protein family, it may be appropriate to regard targets sharing a common Pfam domain or molecular function annotation as replicates, and libraries can be enriched for broad activity within this category, while removing both highly-selective compounds, and promiscuous binders active against a large fraction of the screened targets.

We demonstrate that these data contain a large number of novel active targets for FDA approved drugs, a large number of novel compounds active against known drug targets, and a large set of novel compound-target pairs with no evidence of druggability by FDA approved drugs. By quantifying the rate of agreement between replicated pairs of compound-target activity outcomes, we estimate that less than half of these novel outcomes are due to experimental and data curation errors, and therefore may represent a valuable resource for further drug discovery efforts.

Additionally, we use the statistical model mentioned above to score all highly screened active compounds in PubChem BioAssay by their probability of being promiscuous binders given the available data, and assess the ability of PAINS and aggregators to identify the most experimentally promiscuous compounds. We find that both methods offer mutually complementary strengths at identifying different sets of promiscuous binders, and we also report 1157 compounds with a greater than 50% chance of being promiscuous, that were not included among the sets of known PAINS or aggregators we used for our analysis. We provide the promiscuity probability values for all highly screened active compounds, as well as the source code and results for these analyses as a reference to readers, with the hope that they will contribute to the discovery of medically and biologically useful small molecules.

5.6 Supporting Information

S1 File. Target selectivity, cluster selectivity, domain selectivity, and promiscuity probability $P(\theta \geq 0.25)$ **for all highly screened active compounds.** This is a zipped Excel readable tab separated text file with PubChem compound ids (cid) for each compound in the first column. Compounds are sorted in order from most promiscuous, to most selective. This also serves as a ranked list of target selectivity in reverse order.

S2 File. List of Pfam domains including median target, cluster, and domain selectivities for FDA approved and non-FDA compounds. This is a zipped Excel readable tab separated text file with Pfam identifiers for each domain in the first column. This is the full data shown in tables 5.4 and 5.5, including non-*H. sapiens* domains. All domains with at least one active compound are included.

S3 File. Potentially novel targets for FDA-approved drugs. This is a zipped Excel readable tab separated text file with PubChem compound ids (cids) for each compound in the first column, and a representative UniProt protein target identifier for each sequence-similar target cluster in the second column. These represent compound-target pairs reported as active in PubChem BioAssay, but not represented among the DrugBank annotated targets list. Several targets had no UniProt translation and include a GenBank GI number instead, prefixed with "gi_".

S4 File. FDA approved drug biclusters. This is a zipped Excel readable tab separated text file with PubChem compound ids (cids) for each compound in the first column, and a bicluster for each compound in the second column corresponding to the drug-target biclusters described in the text.

S5 File. Protein target biclusters. This is a zipped Excel readable tab separated text file with a representative UniProt protein identifier for each sequence-similar target cluster in the first column, and a bicluster for each in the second column corresponding to the drug-target biclusters described in the text. Several targets had no UniProt translation and include a GenBank GI number instead, prefixed with “gi_”.

S6 File. Target-protein network. This is a Gephi readable zipped GML (Graph Modeling Language) formatted file, which contains the target-protein network described in the manuscript. Each node (protein) is labeled with a GenBank GI number and a Molecular Function GO slim term.

S7 File. Fully screened compound vs target cluster binary matrix. This is a zipped Excel readable tab separated text file with PubChem compound ids (cid) for each compound in the first column. The first (header) line contains a unique representative UniProt identifier for each sequence-similar protein target cluster. Six targets had no UniProt translation and include a GenBank GI number instead, prefixed with “gi_”. Zero values represent inactive compound-target activity outcomes, while values of one represent active outcomes.

5.7 Bibliography

[1] Tyler William H Backman and Thomas Girke. bioassayR: Cross-Target Analysis of Small Molecule Bioactivity. *J. Chem. Inf. Model.*, 2016.

[2] Jonathan B Baell and Georgina A Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.*, 2010.

[3] M Bastian, S Heymann, and M Jacomy. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 2009.

[4] Y Cao, A Charisi, L C Cheng, T Jiang, and T Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 2008.

[5] T Cheng, Y Wang, and S H Bryant. Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules. *Bioinformatics*, 2010.

[6] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 2011.

[7] G Csardi and T Nepusz. The igraph software package for complex network research. *InterJournal*, 2006.

[8] V Dan ik, H Carrel, N E Bodycombe, K P Seiler, D Fomina-Yadlin, S T Kubicek, K Hartwell, A F Shamji, B K Wagner, and P A Clemons. Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screening*, 2014.

[9] Sivaraman Dandapani and Lisa A Marcaurelle. Grand Challenge Commentary: Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.*, 2010.

[10] Sean R Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, 2009.

[11] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 2007.

[12] Brian Y Feng, Anton Simeonov, Ajit Jadhav, Kerim Babaoglu, James Inglese, Brian K Shoichet, and Christopher P Austin. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.*, 2007.

[13] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 2012.

[14] Malachi Griffith, Obi L Griffith, Adam C Coffman, James V Weible, Josh F McMichael, Nicholas C Spies, James Koval, Indraniel Das, Matthew B Callaway, James M Eldred, Christopher A Miller, Janakiraman Subramanian, Ramaswamy Govindan, Runjun D Kumar, Ron Bose, Li Ding, Jason R Walker, David E Larson, David J Dooling, Scott M Smith, Timothy J Ley, Elaine R Mardis, and Richard K Wilson. DGIdb: mining the druggable genome. *Nat. Methods*, 2013.

[15] L Han, Y Wang, and S H Bryant. A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics*, 2009.

[16] J Harrow, A Frankish, J M Gonzalez, E Tapanari, M Diekhans, F Kokocinski, B L Aken, D Barrell, A Zadissa, S Searle, I Barnes, A Bignell, V Boychenko, T Hunt, M Kay, G Mukherjee, J Rajan, G Despacio-Reyes, G Saunders, C Steward, R Harte, M Lin, C Howald, A Tanzer, T Derrien, J Chrast, N Walters, S Balasubramanian, B Pei, M Tress, J M Rodriguez, I Ezkurdia, J van Baren, M Brent, D Haussler, M Kellis, A Valencia, A Reymond, M Gerstein, R Guigo, and T J Hubbard. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, 2012.

[17] Maria Hauser, Christian E Mayer, and Johannes Söding. kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinf.*, 2013.

[18] M Hewett. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.*, 2002.

[19] E A Howe, A de Souza, D L Lahr, S Chatwin, P Montgomery, B R Alexander, D-T Nguyen, Y Cruz, D A Stonich, G Walzer, J T Rose, S C Picard, Z Liu, J N Rose, X Xiang, J Asiedu, D Durkin, J Levine, J J Yang, S C Schurer, J C Braisted, N Southall, M R Southern, T D Y Chung, S Brudz, C Tanega, S L Schreiber, J A Bittker, R Guha, and P A Clemons. BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.*, 2015.

[20] Ye Hu and Jürgen Bajorath. High-resolution view of compound promiscuity. *F1000Res*, 2013.

[21] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.*, 2012.

[22] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, 2014.

[23] Xavier Jalencas and Jordi Mestres. On the origins of drug polypharmacology. *Med. Chem. Commun.*, 2012.

[24] Swarit Jasial, Ye Hu, and Jürgen Bajorath. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS ONE*, 2016.

[25] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 2007.

[26] Terry Kenakin. Functional selectivity and biased receptor signaling. *J Pharmacol Exp Ther*, 2011.

[27] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem Substance and Compound databases. *Nucleic Acids Res.*, 2016.

[28] Zachary A Knight, Henry Lin, and Kevan M Shokat. Targeting the cancer kinome through polypharmacology. *Nature reviews. Cancer*, 2010.

[29] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, 2010.

[30] Runjun D Kumar, Li-Wei Chang, Matthew J Ellis, and Ron Bose. Prioritizing Potentially Druggable Mutations with dGene: An Annotation Tool for Cancer Genome Sequencing Data. *PLoS ONE*, 2013.

[31] J Laherrère and D Sornette. Stretched exponential distributions in nature and economy: “fat tails” with characteristic scales. *Eur. Phys. J. B*, 1998.

[32] Joseph Lehár, Andrew S Krueger, William Avery, Adrian M Heilbut, Lisa M Johansen, E Roydon Price, Richard J Rickles, Glenn F Short III, Jane E Staunton, Xiaowei Jin, Margaret S Lee, Grant R Zimmermann, and Alexis A Borisy. Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Rev. Drug Discovery*, 2009.

[33] C A Lipinski, F Lombardo, B W Dominy, and P J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 2001.

[34] Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Côté, Brian K Shoichet, and Laszlo Urban. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 2012.

[35] Nathalie Malo, James A Hanley, Sonia Cerquozzi, Jerry Pelletier, and Robert Nadon. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, 2006.

[36] Susan L McGovern, Brian T Helfand, Brian Feng, and Brian K Shoichet. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.*, 2003.

[37] Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, Amaia Sangrador-Vegas, Maxim Scheremetjew, Claudia Rato, Siew-Yit Yong, Alex Bateman, Marco Punta, Teresa K Attwood, Christian J A Sigrist, Nicole Redaschi, Catherine Rivoire, Ioannis Xenarios, Daniel Kahn, Dominique Guyot, Peer Bork, Ivica Letunic, Julian Gough, Matt Oates, Daniel Haft, Hongzhan Huang, Darren A Natale, Cathy H Wu, Christine Orengo, Ian Sillitoe, Huaiyu Mi, Paul D Thomas, and Robert D Finn. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, 2015.

[38] A A Moya-Garcia and J A G Ranea. Insights into polypharmacology from drug-domain associations. *Bioinformatics*, 2013.

[39] Patrick Y Muller and Mark N Milton. The determination and interpretation of the therapeutic index in drug development. *Nat. Rev. Drug Discovery*, 2012.

[40] K D Pruitt, T Tatusova, G R Brown, and D R Maglott. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, 2011.

[41] M Punta, P C Coggill, R Y Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, K Forslund, G Ceric, J Clements, A Heger, L Holm, E L L Sonnhammer, S R Eddy, A Bateman, and R D Finn. The Pfam protein families database. *Nucleic Acids Res.*, 2011.

[42] Simon Saubern, Rajarshi Guha, and Jonathan B Baell. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Molecular Informatics*, 2011.

[43] U Schmidt, S Struck, B Gruening, J Hossbach, I S Jaeger, R Parol, U Lindequist, E Teuscher, and R Preissner. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.*, 2009.

[44] Jean YJ Wang Shun J Lee. Exploiting the promiscuity of imatinib. *Journal of Biology*, 2009.

[45] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 2015.

[46] Miranda van Uitert, Wouter Meuleman, and Lodewyk Wessels. Biclustering Sparse Binary Genomic Data. *J. Comput. Biol.*, 2008.

[47] J D Urban, W P Clarke, M von Zastrow, D E Nichols, B Kobilka, H Weinstein, J A Javitch, B L Roth, A Christopoulos, P M Sexton, K J Miller, M Spedding, and R B Mailman. Functional Selectivity and Classical Concepts of Quantitative Pharmacology. *J Pharmacol Exp Ther*, 2006.

[48] Ubbo Visser, Saminda Abeyruwan, Uma Vempati, Robin P Smith, Vance Lemmon, and Stephan C Schürer. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinf.*, 2011.

[49] Y Wang, J Xiao, T O Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, B A Shoemaker, E Bolton, A Gindulyte, and S H Bryant. PubChem's BioAssay Database. *Nucleic Acids Res.*, 2011.

[50] Yanli Wang, Tugba Suzek, Jian Zhang, Jiyao Wang, Siqian He, Tiejun Cheng, Benjamin A Shoemaker, Asta Gindulyte, and Stephen H Bryant. PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, 2013.

[51] Anne Mai Wassermann, Lisa Peltason, and Jürgen Bajorath. Computational Analysis of Multi-target Structure-Activity Relationships to Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem*, 2010.

[52] D S Wishart, C Knox, A C Guo, D Cheng, S Shrivastava, D Tzur, B Gautam, and M Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 2007.

[53] S Frank Yan, Frederick J King, Yun He, Jeremy S Caldwell, and Yingyao Zhou. Learning from the Data: Mining of Large High-Throughput Screening Databases. *J. Chem. Inf. Model.*, 2006.

[54] M A Yıldırım, K I Goh, M E Cusick, and A L Barabási. Drug—target network. *Nature*, 2007.

[55] Elias Zerhouni. The NIH Roadmap. *Science*, 2003.

[56] Jintao Zhang, Gerald H Lushington, and Jun Huan. Characterizing the Diversity and Biological Relevance of the MLPCN Assay Manifold and Screening Set. *J. Chem. Inf. Model.*, 2011.

Chapter 6

Supporting Information for Large-scale Bioactivity Analysis of the Druggable Proteome

6.1 Target Selectivity Distribution

	Median Selectivity			Mean Selectivity			Trimmed Mean Selectivity		
Compounds	Target	Cluster	Domain	Target	Cluster	Domain	Target	Cluster	Domain
FDA Approved	7	6	4	13.7 ± 19.6	10.6 ± 14.3	5.8 ± 6.5	6.9 ± 5.1	6.4 ± 5.0	4.9 ± 5.2
Non-FDA	2	2	2	4.1 ± 5.8	4.0 ± 5.4	3.5 ± 4.1	3.5 ± 3.5	3.4 ± 3.4	3.2 ± 3.2

Table 6.1: **Distribution of active proteins per compound.** Selectivity is quantified by three methods which cluster protein targets across increasingly large evolutionary distances, as described in the text. Mean values include a standard deviation after the ± symbol. Trimmed mean represents the mean of compounds with values of 20 or fewer. We include a trimmed mean, as mean selectivities are highly skewed by a small number of highly promiscuous compounds, whose selectivity counts reflect the screening data volume rather than the bioactivity profiles.

6.2 Fully Screened Sub-Matrix

As the assay participation for individual compounds is heavily biased, with both very highly screened and infrequently screened compounds (Table 5.1 and Fig 6.1), we sought to identify the largest fully screened sub-matrix of compounds and targets within these data. The size of this sub-matrix quantifies the magnitude of screening bias towards a specific region of the compound-target space, and can also serve as a computational resource for bioactivity analysis methods which cannot accommodate a sparse matrix. Identifying this sub-matrix is a complex nonlinear optimization problem, and simply taking the combination of the most highly screened compounds and targets will not reliably identify the largest bicluster. We chose the BicBin algorithm (set to find non-sparse clusters) for practical reasons, as it scales to large matricies, and finds top-scoring clusters first [16]. As such, we used the BicBin biclustering algorithm to identify the largest fully screened sub-matrix with at least one active outcome in each row and column within the clustered compound-target matrix, as described in the "Fully Screened Sub-Matrix Methods" section of this document. The resulting fully screened matrix had dimensions of 65204 compounds by 260 target clusters, where 23 of the compounds are FDA approved drugs. This number of compounds represents the size of the intersection among the large compound libraries used across many large-scale screening experiments. For example, at the time of writing, PubChem BioAssay contains 128 protein target assays deposited by the NIH Chemical Genomics Center (NCGC) high-throughput screening center that each contain over 70000 compounds. This matrix had an overall activity density (fraction of active scores) of 0.94%, which is very close to the fraction of actives in the full set of PubChem BioAssay protein target data.

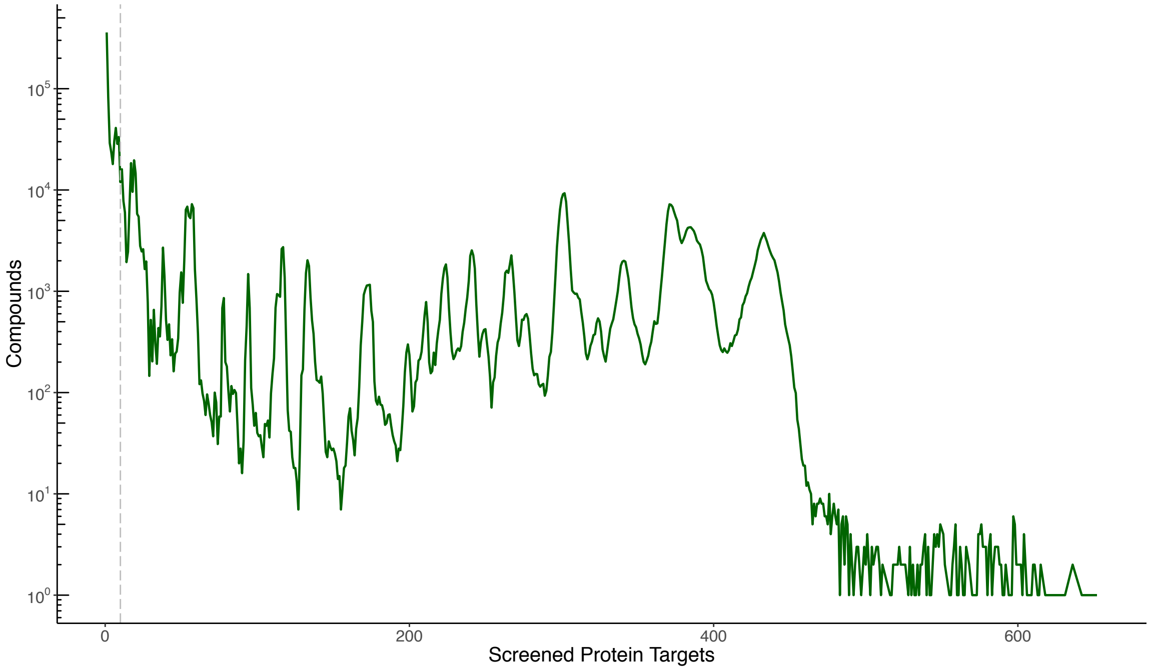


Figure 6.1: **Distribution of distinct protein target assay participation.** Data is included from all assay experiments in PubChem BioAssay annotated with one or more clearly defined protein targets, and reporting an active score for at least one small molecule. The dashed vertical line is drawn at 10 targets, which is the minimum value we categorize in this study as a “highly screened” compound.

This sub-matrix is useful as a representative data set for many data mining techniques. For example, patterns of target selectivity, and compound-target network connectivity can be investigated without introducing bias from varying compound assay participation as is present in the full PubChem BioAssay data. This reference can also be used to train and cross-validate machine learning and imputation methods, in order to assess there performance at imputing missing values from a sparse bioactivity matrix. Lastly, it can be used as a reference for designing custom reduced dimensionality high throughput screening bioactivity fingerprints (HTSFPs), *e.g.* with Principal Component Analysis that still encode a defined portion of the overall variation in bioactivity profiles present in the larger set of public bioactivity data. This matrix is provided as a downloadable resource in Supporting Information, however the full (sparse) bioactivity matrix is used for all analysis in this study, in order to make use of as much relevant data as possible.

6.2.1 Fully Screened Sub-Matrix Methods

We used the BicBin biclustering algorithm to identify the largest fully screened sub-matrix from the clustered compound-target matrix described in the main text, where all values are represented as conclusively active or inactive. We then removed all rows and columns without at least one active experimental outcome. The options $\alpha = 0.5$ and $\beta = 0.8$ were used to bias the result towards including more protein targets, such that the resulting bicluster would include multiple target clusters, instead of summarizing only one or a small number of extremely large assays. We ran the BicBin algorithm enough times to reliably converge on a stable best-scoring bicluster.

6.3 Error Rate Estimate

For those compound-target pairs screened 2-4 times, the quantity of agreement among replicates is shown in Table 6.2. We quantify the number of times replicated sets are all active, all inactive, or have varying levels of disagreement. The error rate in these data was estimated from the pair data (compound-target pairs screened in exactly two assays) by solving the system of equations 6.1, 6.2, and 6.3 below. As the results can be represented by three independent equations, we can solve for three unknowns. In this case we solve algebraically for the number of true positives, true negatives, and error rate. This analysis makes the simplification that the false positive and false negative error rates are identical, as additional information such as a gold standard reference would be necessary to compute these separately.

Times Screened	0 Inactives	1 Inactive	2 Inactives	3 inactives	4 inactives
2	181758	292274	20746238		
3	22740	34868	92323	3158813	
4	8510	11922	20328	31672	654268

Table 6.2: **Disagreement and agreement among compound target pairs screened 2-4 distinct times.** Shown are the number of pairs that had a given number of inactive (vs active) results in these replicates. For example if a given protein target pair was screened in two assays (Times Screened = 2) and both results were inactive it's count would be added to the "2 Inactives" column. If the two assays disagree and one showed active and the other inactive, it is counted in the "1 Inactive" column.

The count of double inactives shown in Equation 6.1 represents the sum of two truly inactive results, plus the number of double false negatives. Here p is the number of true positives, n is the number of true negatives, e is the error rate, I_1 is the set of result pairs where the first was inactive, and I_2 is the set were the second is inactive. I_1^{c} is the set of result pairs where the first was active, and I_2^{c} is the set were the second is active. The order of I_1 and I_2 is irrelevant. The count of single inactives shown in Equation 6.2 represents the sum of inactive pairs with one false positive, and active pairs with one false negative. The count of zero inactives shown in Equation

6.3 represents the sum of two truly active results, plus the number of double false positives.

$$\overbrace{|I_1 \cap I_2|}^{\text{two inactives count}} = \overbrace{n(1-e^2)}^{\text{double true negatives}} + \overbrace{pe^2}^{\text{double false negatives}} \tag{6.1}$$

$$\overbrace{|I_1 \triangle I_2|}^{\text{one inactive count}} = \overbrace{2ne}^{\text{one false positive}} + \overbrace{2pe}^{\text{one false negative}} \tag{6.2}$$

$$\overbrace{|I_1^c \cap I_2^c|}^{\text{zero inactives count}} = \overbrace{p(1-e^2)}^{\text{double true positives}} + \overbrace{ne^2}^{\text{double false positives}} \tag{6.3}$$

Solving for e based on the data in the first row of Table 6.2 yields an estimated error rate of approximately $e = 0.00698$, or roughly 0.7% with approximately $p =$181k true positives, and approximately $n =$20.7M true negatives. If the false positive and false negative rates were identical, this would represent a fraction of true active compound-target outcomes of about 0.86%, which is somewhat lower than the 1.3% of values which are active in the entire PubChem BioAssay protein target data we analyze in this study. If we additionally make the approximation that the fraction of true positives in this replicated data is close to that in the larger non-replicated set, we estimate that the overall fraction of active results which are true positives is roughly $100 * 0.86/1.3 = 66\%$, ignoring the negligible false negatives. This error rate is low enough that only a small portion (roughly 0.56%) of the double positive (0 inactives) results are likely to be double false positives rather than true positives, while most of the single inactive results among replicated pairs are a result of false positives.

6.4 Target Selectivity by Molecular Size

To investigate the possibility of a relationship between target selectivity and molecular size, we quantified the distribution of molecular sizes for compounds with different numbers of distinct sequence active targets, *i.e.* cluster selectivity. The definition of molecular size used here is the quantity of non-hydrogen atoms. We analyzed these data to identify both general trends of selectivity in relation to molecular size, as well as look for distinct patterns among the largest and smallest compounds. All highly screened compounds with between 1 and 10 active targets are shown in Fig 6.2. Many of the very largest compounds in PubChem BioAssay with over 300 heavy atoms were screened in a large number of assays but reported inconclusive results in most or all, and were excluded here. This suggests that there may be technical limitations to these assay methodologies with regard to extremely large compounds. Non-FDA compounds have a slightly higher mean molecular size than FDA approved drugs (Table 6.3 and Fig 6.2). While the overall distribution of molecular size is roughly the same across compounds with differing numbers of active targets, the largest and smallest compounds show greater target selectivity (fewer targets) than non-FDA compounds, as can be seen in the tail lengths of Fig 6.2. While these compounds vary in size from 1 to 302 heavy atoms, all compounds with a size under 3 or over 190 (a total of 4 and 8 respectively) have 3 or fewer active targets. Compounds with greater than 10 active protein targets show approximately the same distribution of size shown here for 7-10 active targets, and were excluded from the plot due to space considerations.

Compounds	Mean±SD Size	Size Range
FDA Approved	23.0 ± 9.3	$4 - 87$
Other	26.3 ± 7.0	$2 - 302$

Table 6.3: **Variation in molecular size for FDA approved and non-FDA compounds.** FDA approved compounds have a slightly lower mean size than non-FDA compounds. Only highly screened active compounds were included in this calculation.

The target selectivity of the 40 largest highly screened active FDA approved drugs is shown in Table 6.4, along with the common name and therapeutic utility (annotation from DrugBank). A majority of these compounds are either natural products or semi-synthetic drugs derived from natural products. It has been reported in previous literature that approximately half of naturally discovered drugs violate the Lipinski Rule of Five in terms of molecular size, and number of rotatable bonds, despite the fact that effective and bioavailable synthetic drugs rarely violate these rules [9, 13]. As discussed by Ganesan, this may in part be due to naturally evolved synthetic processes that allow large molecules to maintain low hydrophobicity and intermolecular H-bond donating potential, as well as the ability of natural products to mimic endogenous metabolites allowing them to utilize active transport systems [9]. Overall this subset shows activity against a greater number of targets than most FDA approved compounds (median 12.5 vs 7 for all drugs), however there is substantial variation by drug class. Several of the most selective large molecules such as streptomycin and amikacin are natural antibiotics evolved to specifically inhibit prokaryotic ribosomal RNA structures, which are not among the screening targets in our analysis [2, 3]. Many of the least selective large molecules are highly promiscuous cytotoxic compounds (such as Suramin) which have dual utility as cancer chemotherapy agents and as antiparasitic drugs [17, 4, 14].

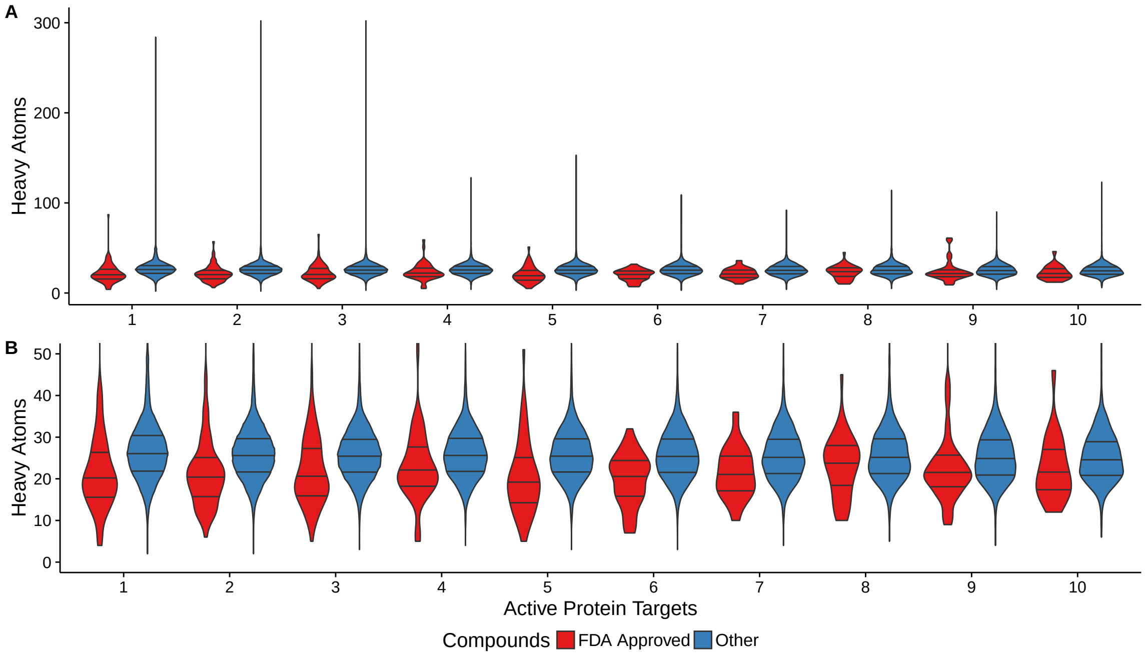


Figure 6.2: **Target selectivity by molecular size.** Violin plot with horizontal lines drawn at the 0.25, 0.5, 0.75 quantiles with tails trimmed to the range of data. Molecule size is quantified here by the number of non-hydrogen (heavy) atoms. (A) Target selectivity vs. molecular size across the entire range (y axis) of variation in these data. (B) Target selectivity vs. molecular size zoomed in on the y-axis to show more detail.

Name	PubChem CID	Size	Active Targets	Tested Targets	Description
Cisatracurium Besylate	62886	87	1	59	muscle relaxant
Suramin	5361	86	22	63	antiparasitic
Amphotericin B	5280965	65	3	71	antifungal
Paclitaxel	36314	62	20	293	antimitotic
Rifabutin	6323490	61	9	228	antibiotic
Vincristine	5978	60	17	57	antimitotic
Vinblastine	241903	59	4	25	antimitotic
Docetaxel	148124	58	9	33	antimitotic
Rifaximin	6436173	57	2	15	antibiotic
Digitoxin	441207	54	28	229	cardiac glycoside
Erythromycin	12560	51	19	330	antibiotic
Posaconazole	147912	51	5	28	antifungal
Atazanavir	148192	51	6	36	antiretroviral protease inhibitor
Ritonavir	392622	50	26	243	antiretroviral protease inhibitor
Itraconazole	55283	49	15	34	antifungal
Lopinavir	92727	46	14	76	antiretroviral protease inhibitor
Rescinnamine	5280954	46	10	260	antihypertensive
Tubocurarine	6000	45	2	61	neuromuscular blocker
Lercanidipine	65866	45	3	39	antihypertensive
Indinavir	5362440	45	10	10	antiretroviral protease inhibitor
Reserpine	5770	44	28	339	antipsychotic
Ergotamine	8223	43	38	42	vasoconstrictor
Dihydroergotamine	10531	43	31	88	vasoconstrictor
Bromocriptine	31101	43	37	77	dopamine agonist
Irinotecan	60838	43	11	22	antimitotic
Nonoxynol-9	72385	43	1	44	surfactant
Benzonatate	7699	42	2	322	antitussive
Deserpidine	8550	42	11	24	antihypertensive
Etoposide	36462	42	22	273	antimitotic
Zafirlukast	5717	41	42	229	leukotriene receptor antagonist
Ouabain	439501	41	19	298	cardiac glycoside
Montelukast	5281040	41	25	28	leukotriene receptor antagonist
Hexafluronium	9434	40	3	14	neuromuscular blocker
Streptomycin	19649	40	1	34	antibiotic
Amikacin	37768	40	1	43	antibiotic
Nelfinavir	64143	40	27	321	antiretroviral protease inhibitor
Lapatinib	208908	40	17	166	tyrosine kinase inhibitor
Deferoxamine	2973	39	5	182	chelating agent
Doxorubicin	31703	39	30	43	antitumor antibiotic
Telmisartan	65999	39	15	289	antihypertensive

Table 6.4: **Target selectivity of the 40 largest highly screened active FDA Approved Drugs.** Sorted by decreasing size (non-hydrogen atom count). The description column represents one common clinical use of each compound, but may not represent its only known therapeutic utility.

6.5 Stretched Exponential Selectivity Distribution

As a curvature can be seen in the semi-log plot space for each of the target selectivity distributions shown in S6 Fig, these data have a regular pattern that cannot be described fully by an exponential probability mass distribution. To model the distribution of target selectivity we fit the distinct sequence target data to three probability density function models, the exponential (equation 6.4), power law (equation 6.5), and stretched exponential function (equation 6.6) where x represents the number of active targets, and each function includes two fit parameters. While the number of active targets for each compound is a discrete value suggesting a probability mass function, here we fit a continuous function to approximate it. The exponential and power law functions fit with roughly opposite sign residuals ($R^2 = 0.99131$ and $R^2 = 0.97916$ respectively), and the stretched exponential, which is a composite of the other two, fit the data closely ($R^2 = 0.99912$) as shown in Fig 6.3.

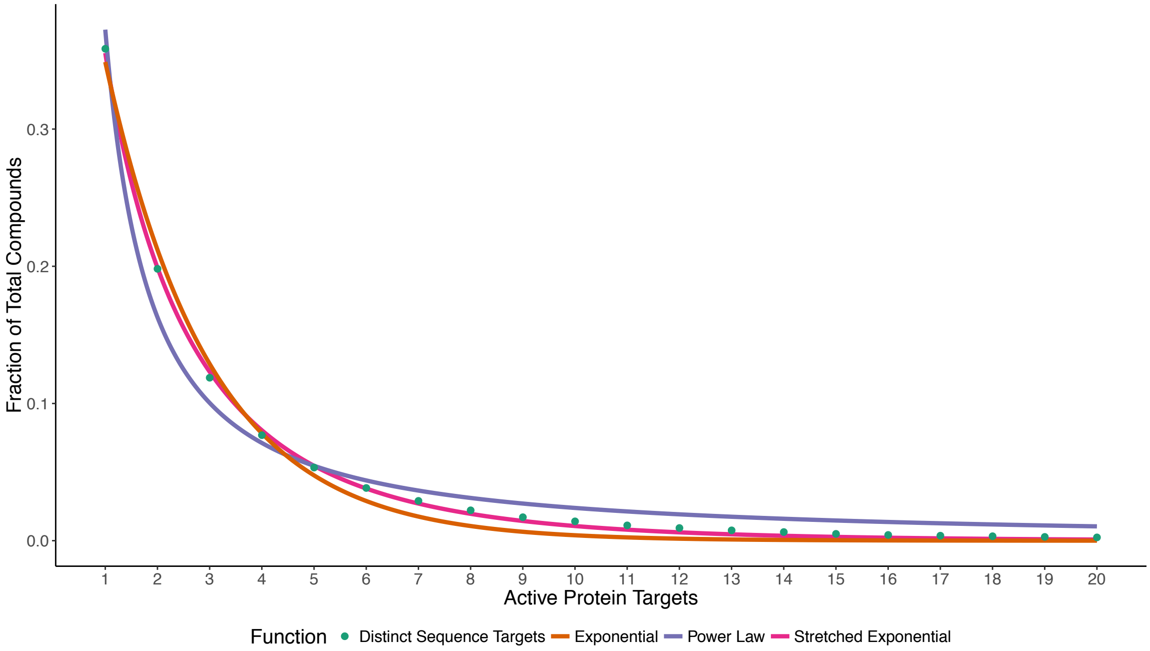


Figure 6.3: **Selectivity function fit.** The distribution of distinct sequence targets for non-FDA approved compounds, along with best fit lines using two-parameter versions of the exponential, power law, and stretched exponential functions.

$$P(x) = fe^{-(\frac{x}{g})} \tag{6.4}$$

$$P(x) = ax^b \tag{6.5}$$

$$P(x) = e^{-(\frac{x}{x_0})^c} \tag{6.6}$$

The stretched exponential fit the non-FDA cluster selectivity distribution for 1-20 targets with best fit parameters $c = 0.6423$ and $x_0 = 0.9487$. Fitting the FDA approved drug data to the stretched exponential (equation 6.6, $R^2 = 0.9033$) yielded best fit parameters $c = 0.2445$ and $x_0 = 0.06681$. The FDA approved distribution was fit slightly better by an exponential distribution, and exhibits much less curvature in semi-log space. The fact that the exponential function fits relatively well suggests that the distribution of active targets for each compound approximately follows an exponential underlying process. That is, each compound-target interaction can be seen as an independent event of roughly equal probability. However, if these were singular independent outcomes with the same probability we would expect the exponential function to fit better than the stretched exponential, but as shown that is not the case for the non-FDA compounds.

Stretched exponential probability distributions have been documented in a large number of natural and artificial processes, many of which have been shown to follow this probability distribution due to an underlying multiplicative process, where a series of chained events, each with different probabilities is involved in the resulting outcome quantified [11, 8, 15, 10]. For example, they can be observed in the distribution of binding times for single molecule enzyme-substrate combinations, and protein structural relaxation times [5, 6, 1, 12, 7]. This lends a possible physical interpretation to the two fit parameters in these processes. The value of the exponent, c describes the curvature in semilog space, and arises from the underlying multiplicity of the process. In a stretched exponential, the exponent c is always smaller than one, with a value inverse to the number of generations, $m = \frac{1}{c}$ in the underlying multiplicative cascade [11]. Further investigation will be necessary to elucidate the nature of the possible key underlying multiplicative steps involved in compound-target interactions suggested by this analysis.

6.5.1 Stretched Exponential Selectivity Distribution Methods

Numeric fit was computed with the *nls* (Nonlinear Least Squares) function in the R programming language. Each function included two fit parameters.

6.6 Additional Figures

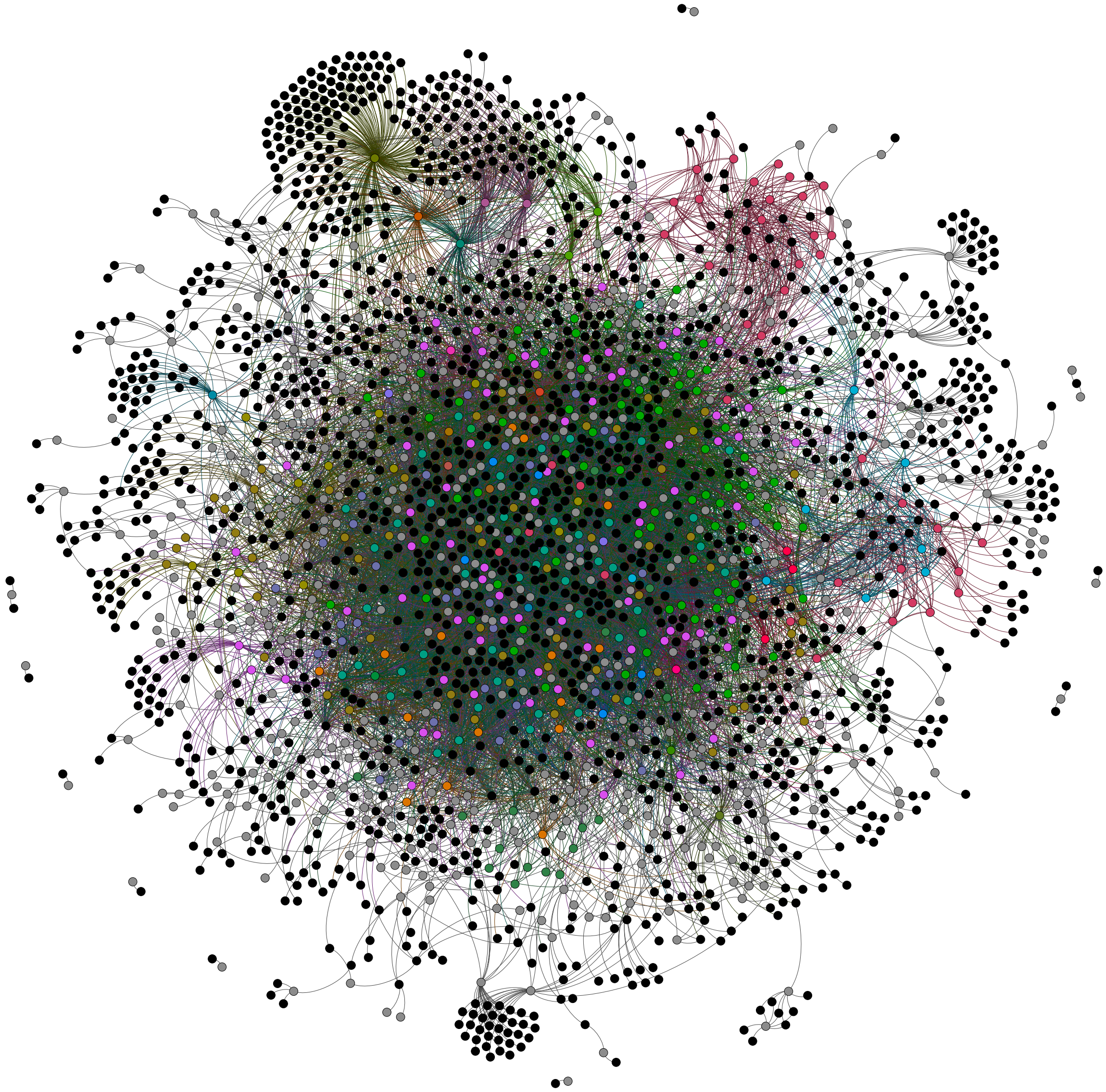


Figure 6.4: **Drug-Target (DT) bipartite network biclusters.** Protein targets are shown in black, with FDA approved drugs shown in color, based on their bioactivity bicluster. Unclustered compounds are shown in grey. No color key is provided, as some colors were reused in order to visualize a large number of biclusters. Node position is based on connectivity, with the same positions as in Fig 6.5.

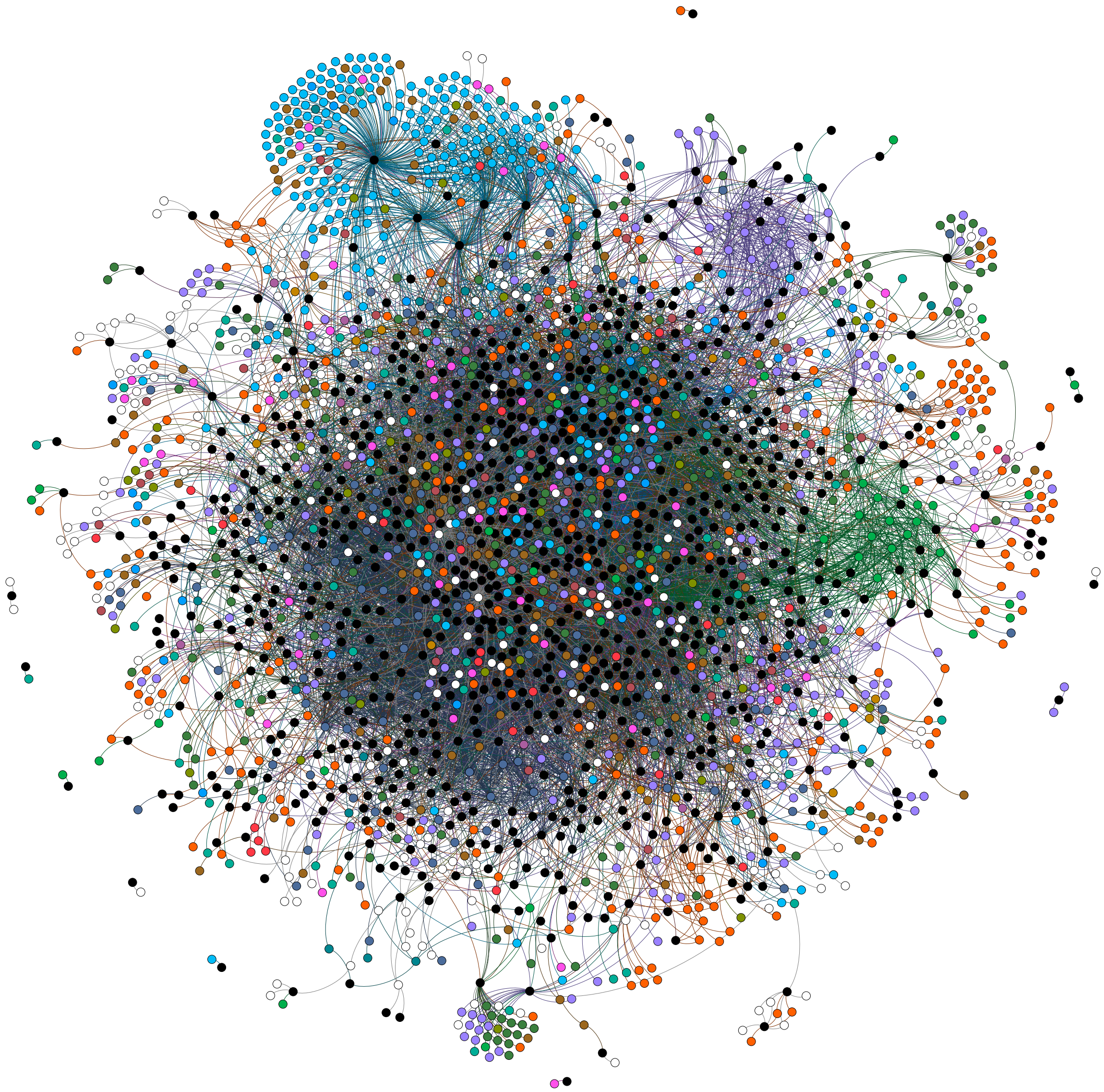


Figure 6.5: **Drug-Target (DT) bipartite network Gene Ontology (GO).** FDA approved drugs are shown in black, with protein targets show in color based on the most specific Molecular Function GO Slim term for each target. Unannotated targets are shown in white. No color key is provided, as some colors were reused in order to visualize a large number of GO terms. Node position is based on connectivity, with the same positions as in Fig 6.4.

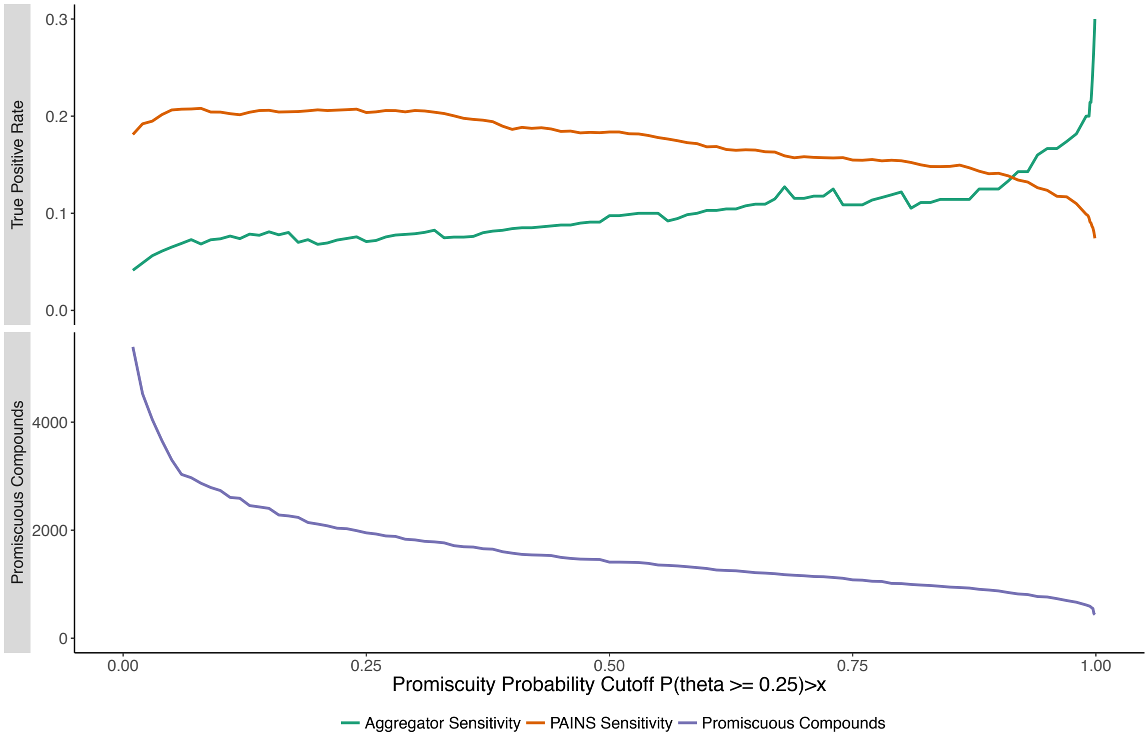


Figure 6.6: **Sensitivity of PAINS and aggregators vs promiscuity probability cutoff.** The top panel shows the sensitivity (true positive rate) of PAINS and aggregators to categorize promiscuous compounds throughout a range of promiscuity probability cutoffs $P(\theta \geq 0.25) > x$ over the range $x = [0.01, 0.9999]$. The bottom panel shows the number of promiscuous compounds at each cutoff value.

6.7 Bibliography

[1] R H Austin, K W Beeson, and L Eisenstein. Dynamics of ligand binding to myoglobin. *Biochemistry*, 1975.

[2] Erik C Böttger, Burkhard Springer, Therdsak Prammananan, Yishak Kidan, and Peter Sander. Structural basis for selectivity and toxicity of ribosomal antibiotics. *EMBO reports*, 2001.

[3] A P Carter, W M Clemons, D E Brodersen, R J Morgan-Warren, B T Wimberly, and V Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, 2000.

[4] Nancy A Dawson, Michael R Cooper, William D Figg, Donna J Headlee, Alain Thibault, Raymond C Bergan, Seth M Steinberg, Edward A Sausville, Charles E Myers, and Oliver Sartor. Antitumor activity of suramin in hormone-refractory prostate cancer controlling for hydrocortisone treatment and flutamide withdrawal as potentially confounding variables. *Cancer*, 1995.

[5] P W Fenimore, H Frauenfelder, B H McMahon, and R D Young. Proteins are paradigms of stochastic complexity. *Physica A: Statistical Mechanics and its Applications*, 2005.

[6] Ophir Flomenbom, Kelly Velonia, Davey Loos, Sadahiro Masuo, Mircea Cotlet, Yves Engelborghs, Johan Hofkens, Alan E Rowan, Roeland J M Nolte, Mark Van der Auweraer, Frans C de Schryver, and Joseph Klafter. Stretched exponential decay and correlations in the catalytic activity of fluctuating single lipase molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 2005.

[7] Hans Frauenfelder, Peter G Wolynes, and Robert H Austin. Biological Physics. *Rev. Mod. Phys.*, 1999.

[8] U Frisch and D Sornette. Extreme Deviations and Applications. *J. Phys. I France*, 1997.

[9] A Ganesan. The impact of natural products upon modern drug discovery. *Current Opinion in Chemical Biology*, 2008.

[10] J Klafter and M F Shlesinger. On the relationship among three theories of relaxation in disordered systems. *Proc. Natl. Acad. Sci. U.S.A.*, 1986.

[11] J Laherrère and D Sornette. Stretched exponential distributions in nature and economy: “fat tails” with characteristic scales. *Eur. Phys. J. B*, 1998.

[12] David G Lambright, Sriram Balasubramanian, and Steven G Boxer. Protein relaxation dynamics in human myoglobin. *Chemical Physics*, 1991.

[13] C A Lipinski, F Lombardo, B W Dominy, and P J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 2001.

[14] Martin C Michel and Roland Seifert. Selectivity of pharmacological tools: implications for use in cell physiology. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms. *American Journal of Physiology - Cell Physiology*, 2015.

[15] G G Naumis and G Cocho. The tails of rank-size distributions due to multiplicative processes: from power laws to stretched exponentials and beta-like functions. *New Journal of Physics*, 2007.

[16] Miranda van Uitert, Wouter Meuleman, and Lodewyk Wessels. Biclustering Sparse Binary Genomic Data. *J. Comput. Biol.*, 2008.

[17] T E Voogd, E L Vansterkenburg, J Wilting, and L H Janssen. Recent research on the biological activity of suramin. *Pharmacol Rev*, 1993.

Chapter 7

Receptor-Mediated Leptin Transcytosis: a Model of the Blood-Brain Barrier as a Lipostatic Regulatory Interface

7.1 Background and Introduction

The leptin molecule (coded for by the mouse *ob* gene and it's human orthologue) has been identified as a critical protein hormone in the physiological regulation of energy homeostasis in mammals[31]. This hormone is produced by adipose tissue, and serum leptin levels correlate positively with adipose mass implicating it as a potential "adipostat" signal to the hypothalamus[7]. Decreased blood leptin levels in *ob* knockout mice are correlated with decreased energy expenditure, and increased food consumption[31]. This discovery led to initial enthusiasm amongst researchers and health professionals as a route for the development of novel treatments for obesity and associated metabolic disorders by increasing serum leptin concentration[10]. Unfortunately, it was later discovered that increasing blood leptin levels in both obese humans and rats fails to alter energy homeostasis as leptin levels in the brain and cerebral spinal fluid (CSF) remain roughly unchanged[20]. In animal models of obesity, administration of leptin directly to the central nervous system (CNS) alters energy homeostasis, while administration into the blood fails to do so[25]. This implicates reduced leptin transport across the blood-brain barrier and blood-CSF barrier as a potential factor in leptin resistance, consistent with the identification of a saturable receptor-mediated transport system for leptin[3].

The “lipostasis” hypothesis posits that the hypothalamus regulates energy balance in order to maintain adipose mass within a narrow range. There are two theorized processes by which the “adipostat” signal provided by leptin could be interrupted and cause the hypothalamus to defend a higher adipose mass: (1) inflammation in hypothalamus leading to decreased leptin sensitivity, and (2) deficient transport of leptin from the blood to the central nervous system [30, 20]. Evidence that supports the existence of such a mechanism in humans includes: increasing caloric expenditure (exercise) leads to a corresponding increase in hunger, adipose mass loss due to forced calorie restriction partially reverses if restriction is ceased, and adipose mass gain partially reverses if forced overfeeding ceases[13, 26, 6]. The term “leptin resistance” refers to a pathological condition which leads to the hypothalamus defending a higher adipose mass.

The blood-brain and blood-CSF barriers typically block the transport of large molecules such as leptin from entering the central nervous system. Rat brain capillaries at the blood-brain barrier exhibit high expression of a shortened form of the leptin receptor (ObRa), which (unlike the long form ObRb) have not been related to any intracellular signal transduction pathways but has been demonstrated to be capable of unidirectional leptin transport *in vitro*[4, 9]. This suggests a mechanism by which leptin could gain access to the brain. Other receptor isoforms also exist, but have unknown function[27].

Several factors have been demonstrated as modulators of leptin transport at the blood brain and and blood-CSF barriers, and consequently as potentially important factors in both the pathophysiology and treatment of leptin resistance. These include serum triglycerides, lipopolysaccharides, and soluble leptin receptors which decrease transport, and α_1 adrenergics which increase leptin transport[1, 2, 23].

Here I first develop a quasi-steady-state mathematical model of the ObRa leptin transport system which predicts saturable behavior consistent with experimental data. I then extend the model to incorporate receptor population and accommodate time varying behavior in response to external factors. Lastly, I propose a series of experiments to measure unknown rate constants in this extended model, which represents a "high throughput hypothesis" incorporating several potential biochemical mechanisms by which leptin transport could be impaired. Fitting this model computationally to data from a carefully designed series of *in vitro* experiments comparing cells transporting (or endocytosing) leptin by the ObRa receptor with and without known factors which impair leptin transport will measure the unknown rate constants for both systems, and help to refine the hypothesis of how the transport mechanism is being modified in cases of pathological leptin resistance.

Similar experiments in the past have proven capable of independently measuring the unknown rate constants of several natural receptor mediated endocytosis mechanisms. Currently the best characterized receptor-mediated transport system is the cellular uptake of epidermal growth factor (EGF) by it's receptor (EGFR). Researchers were able to accurately measure the unknown rate constants in this mechanism by fitting it's dynamic *in vitro* behavior to mathematical models[28, 29]. To the best of my knowledge this technique has not yet been extended to elucidate the biochemical mechanism by which various factors influence receptor mediated transcytosis.

7.1.1 Transport Physiology

Six distinct isoforms (splice variants) of the mouse and human leptin receptor (ObR gene) have been identified. All six of these isoforms are capable of binding to the leptin molecule (Ob gene) and localizing on the exterior of cellular membranes. The long isoform (ObRb) is preferentially expressed in the hypothalamus and is capable of intracellular signal transduction via MAPK and Jak-Stat signaling cascades, where it inhibits appetite and alters other factors related to energy homeostasis [15]. Five short isoforms have been identified (ObRa, ObRc, ObRd, ObRe, and r-ObRf) which lack motifs associated with signal transduction[27]. ObRa has been identified as playing

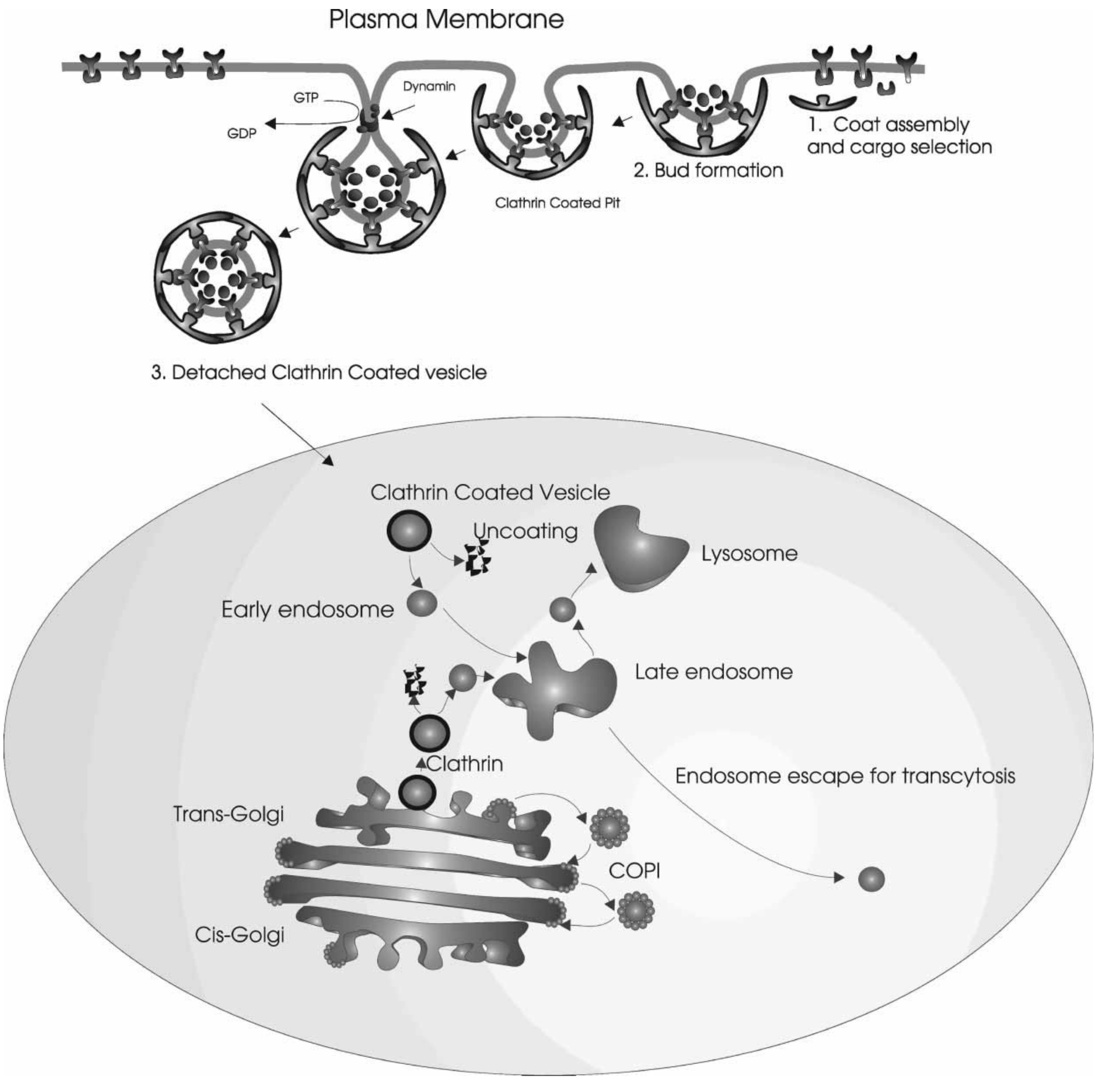


Figure 7.1: **Illustration of a generalized clarathrin-dependent receptor mediated transcytosis mechanism.** Figure reprinted from Smith and Gumbleton 2006 with permission from the publisher[21].

an essential role in transport and lysosomal degradation of leptin with internalization mediated by a clathrin-coated pit endocytosis mechanism[24].

Unidirectional transport of large protein molecules by clathrin-coated pit mechanisms at the blood-brain and blood-CSF barriers involve expression of the receptor on the blood plasma side of the endothelial membrane which consists of polarized cells. Transport receptors diffuse along the surface, and are capable of binding to ligands freely diffusing in plasma. Once bound to a ligand, the bound complex may either dissociate or locate to a clathrin-coated pit (which make up a portion of the membrane) for endocytosis. Some receptors undergo a conformational change after binding which increases the probability of coated-pit localization, but this does not seem to occur for ObRa. Coated pits regularly form vesicles and internalize, and any receptors bound to them at this time will become entrapped and internalized in the resulting vesicle. This vesicle may also contain unbound receptors and other molecules from blood plasma.

After internalization, clathrin will dissociate from the vesicle transforming it into an early endosome which experiences a drop in pH due to ion pumps on its membrane. This drop in pH causes the receptors and ligands to dissociate and sort by an unknown mechanism inside what is now a late endosome. After this sorting occurs, the late endosome splits in two, where one half contains mostly receptors, and other half mostly ligand. A fraction of both the receptor and ligand portions of the late endosome merge with lysosomes resulting in lysosomal degradation. The fraction of receptors that escape degradation are recycled back to the cell surface while (in the case of transcytosis) the remaining ligand rich portions of the late endosome are released on the opposite side of the cell (see figure 7.1).

7.2 Methods (Transport Kinetics)

7.2.1 Quasi-steady-state Transport Rate

Leptin transport is hypothesized to occur via a clarathrin-dependent receptor mediated transport mechanism with general steps of binding and dissociation on the cell surface, internalization, and both transport and lysosomal degradation of internalized leptin. Such a mechanism is represented as follows:

$$Lep + ObRa \quad \overbrace{\underbrace{\underset{k_1}{\overset{k_{-1}}{\leftrightharpoons}}}_{\text{Binding}}}^{\text{Dissociation}} \quad LepObRa \quad \underbrace{\underset{k_2}{\rightarrow}}_{\text{Internalization}} \quad Lep_i \quad \overbrace{\underset{k_3}{\rightarrow}}^{\text{Transport and Degradation}} \quad Lep_d + Lep_t \tag{7.1}$$

Where $[Lep]$ is blood leptin concentration, $[ObRa]$ is the number of unbound surface transport receptors on the blood side of the barrier, $[LepObRa]$ is the number of leptin-receptor complexes on the surface, $[Lep_i]$ is the number of internalized leptin molecules within the cell(s), $[Lep_d]$ is the number of leptin molecules delivered for lysosomal degradation, and $[Lep_t]$ is the number of leptin molecules transported across the membrane.

Applying conservation of mass to the leptin molecule for each step in this mechanism yields the following differential equations for leptin concentration in each phase:

$$\frac{d[LepObRa]}{dt} = \overbrace{k_1[Lep][ObRa]}^{\text{Binding}} - \overbrace{(k_2 + k_{-1})[LepObRa]}^{\text{Dissociation and Internalization}}$$

$$\frac{d[Lep_i]}{dt} = \overbrace{k_2[LepObRa]}^{\text{Internalization}} - \overbrace{k_3[Lep_i]}^{\text{Transport and Degradation}} \tag{7.2}$$

$$\frac{d[Lep_d + Lep_t]}{dt} = \overbrace{k_3[Lep_i]}^{\text{Transport and Degradation}} \tag{7.3}$$

For a quasi-steady-state where blood leptin and receptor population are constant ($d[Lep]/dt = 0$,$d([LepObRa] + [ObRa])/dt = 0$) and the system is at equilibrium ($t >> t_0$) then the bound ligand/receptor complex, and internal leptin levels will also reach constant values:

$$k_1[Lep][ObRa] = (k_2 + k_{-1})[LepObRa] \tag{7.4}$$

$$k_2[LepObRa] = k_3[Lep_i]$$

This simplification allows for the elimination of $[Lep_i]$ from the model:

$$\frac{d[Lep_d + Lep_t]}{dt} = k_2[LepObRa]$$

$$\frac{d[Lep_t]}{dt} = k_2(1 - f_L)[LepObRa] = k_4[LepObRa] \tag{7.5}$$

Where $k_4 \equiv k_2(1 - f_L)$, with f_L being the fraction of leptin sorted for lysosomal degradation (also constant at steady state). Substituting the constant receptor population ($[ObRa]_0 = [LepObRa] + [ObRa]$) into equation (7.4) yields the following:

$$k_1[Lep]([ObRa]_0 - [LepObRa]) = (k_2 + k_{-1})[LepObRa]$$

Which can be solved algebraically for [LepObRa]:

$$[LepObRa] = \frac{[Lep][ObRa]_0}{\frac{k_2+k_{-1}}{k_1} + [Lep]}$$

Substituting into equation (7.5) yields an equation for the rate of transport:

$$\frac{d[Lep_t]}{dt} = k_2(1-f_L)\frac{[Lep][ObRa]_0}{\frac{k_2+k_{-1}}{k_1} + [Lep]} = k_4\frac{[Lep][ObRa]_0}{\frac{k_4/(1-f_L)+k_{-1}}{k_1} + [Lep]}$$

This has the same mathematical form as the Michaelis-Menten equation of enzyme kinetics, which exhibits saturable behavior:

$$v_0 = \frac{v_{max}[Lep]}{k_M + [Lep]} \tag{7.6}$$

Where v_0 is the rate of transport for a given $[Lep]$, v_{max} is the maximum rate of transport at saturation, and k_M is a constant at quasi-steady-state:

$$v_0 \equiv \frac{d[Lep_t]}{dt}$$

$$v_{max} \equiv k_4[ObRa]_0$$

$$k_M \equiv \frac{k_4/(1-f_L)+k_{-1}}{k_1}$$

7.2.2 Diffusion vs Transport Rate

The above model of transport behavior considers the rate of transport relative to concentration of leptin at the blood-brain and blood-CSF barrier surface, but does not

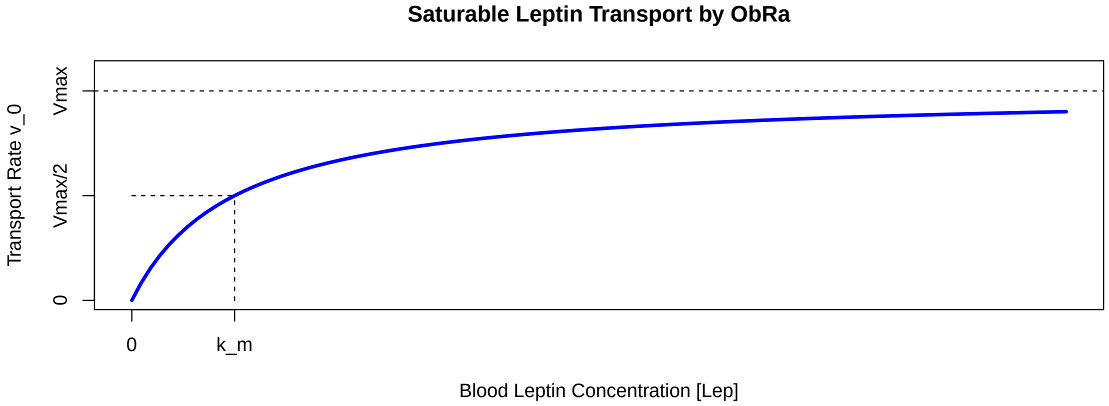


Figure 7.2: **Equation** (7.6)**: Unit-less saturable response of transport rate to serum leptin levels.**

consider the possibility of a leptin concentration gradient forming in the capillaries, which would limit the real world maximum transport rate via diffusion rather than the above calculated v_{max}. For purposes of understanding leptin resistance I am concerned primarily with factors which could suppress real world v_{max} at brain regions of rapid uptake, and therefore inhibit leptin sensitivity within the higher ranges of physiologically relevant leptin levels. Therefore, estimating the relative contribution of diffusion and transcytosis at the blood-CSF barrier in adult rats is representative, as this is the site of most rapid leptin transport under physiological conditions[32]. The Damköhler number (D_a) is a unit-less comparison of the rate of transport (v_{max}) to the rate of leptin diffusion (DC/L).

$$v_{max} \leq 0.2 \frac{ng}{g * min} \frac{1min}{60s} \frac{1g}{1ml} \frac{0.58ml}{75cm^2} = 2.57 * 10^{-5} \frac{ng}{s * cm^2} \tag{7.7}$$

$$D_a = \frac{v_{max} L}{DC} \leq \frac{2.57 * 10^{-5} \frac{ng}{s*cm^2} * 2 * 10^{-4} cm}{0.061 * 10^{-5} \frac{cm^2}{s} * 0.6 \frac{ng}{cm^3}} \approx 0.014 << 1$$

Where v_{max} is the rate of leptin transport at receptor saturation per unit area $(\frac{ng}{s*cm^2})$, L is the characteristic length (length over which a concentration gradient may occur), D the binary diffusion coefficient of leptin in blood, and C is the serum leptin concentration (see figure 7.3). Each of these parameters can occur over a wide range *in vivo*, so an attempt was made to use maximal values for estimating parameters in the numerator, and minimal values in the denominator to err on the side of overestimating the Damköhler number.

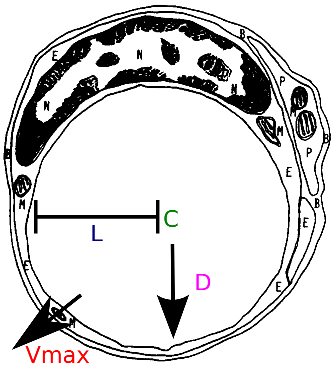


Figure 7.3: **Diagram of diffusion vs transport in the rat brain capillary.** v_{max} represents the maximal rate of transport, D the binary diffusion coefficient of leptin in blood, C the serum leptin concentration, and L the capillary radius. Figure adapted from Hicks et al. 1983 with permission from Elsevier[8].

v_{max} was estimated by taking the measured v_{max} at the site of fastest transport within the "rapid uptake" portions of the brain per *in vivo* rat models which is $v_{max} \leq 0.2 \frac{ng}{g*min}$ into cerbero-spinal fluid[32]. This value was converted to a mass flux (equation 7.7) by using published parameters on the average ratio of cerebrospinal fluid (CSF) to choroid plexus surface area in the adult rat brain[11, 14]. CSF has a density of roughly 1 gram per ml[19]. The intricate geometry of the choroid plexus accounts for the huge surface area relative to such a small fluid volume. Serum leptin concentration C was estimated from the low end of physiological leptin concentrations ($C \geq 0.6\frac{ng}{cm^3}$)[32]. The binary diffusion coefficient of leptin is unknown, but is assumed to be of the same order of magnitude as albumin ($D \approx 0.061 * 10^{-5}\frac{cm^2}{s}$) [5]. Since the estimated Damköhler number is much less than unity, diffusion is unlikely to limit the rate of transport and will be omitted from the model presented herein.

7.2.3 Dynamic Behavior and Receptor Regulation

The leptin transport receptor ObRa occurs on the cell surface at varying population levels. The following reaction represents the mechanism by which surface receptors are hypothesized to undergo synthesis, binding, internalization, surface recycling, and lysosomal degradation:

$$\underbrace{\underset{S}{\rightarrow}}_{\text{Receptor Synthesis}} \qquad ObRa + LepObRa \quad \overbrace{\underbrace{\underset{k_2}{\overset{k_{rec}(1-f_R)}{\leftrightharpoons}}}_{\text{Internalization}}}^{\text{Recycling}} \quad ObRa_i \quad \overbrace{\underset{k_{deg}f_R}{\rightarrow}}^{\text{Degradation}} \quad ObRa_d \tag{7.8}$$

Where S is the rate of receptor synthesis, $[ObRa + LepObRa]$ is the total surface receptor population (unbound and bound), k_2 is the rate of internalization, k_{rec} is the rate of receptor recycling to the cell surface, f_R is the fraction of internal receptors sorted for lysosomal degradation, $[ObRa_i]$ is the population of internal receptors, k_{deg} is the rate of lysosomal degradation, and $[ObRa_d]$ is the quantity of receptors that has undergone degradation.

Applying conservation of mass to the receptor population for each step in this mechanism yields the following differential equations for receptor population in each phase:

$$\begin{aligned}\frac{d[ObRa]}{dt} &= \overbrace{k_{-1}[LepObRa]}^{\text{Dissociation}} - \overbrace{k_1[Lep][ObRa]}^{\text{Binding}} - \overbrace{k_2[ObRa]}^{\text{Internalization}} \\ &\quad + \overbrace{k_{rec}(1-f_R)[ObRa_i]}^{\text{Recycling}} + \overbrace{S}^{\text{Receptor Synthesis}}\end{aligned} \tag{7.9}$$

$$\frac{d[LepObRa]}{dt} = \overbrace{k_1[Lep][ObRa]}^{\text{Binding}} - \overbrace{k_{-1}[LepObRa]}^{\text{Dissociation}} - \overbrace{k_2[LepObRa]}^{\text{Internalization}} \tag{7.10}$$

$$\frac{d[ObRa_i]}{dt} = \overbrace{k_2([LepObRa]+[ObRa])}^{\text{Internalization}} - \overbrace{k_{rec}(1-f_R)[ObRa_i]}^{\text{Recycling}} - \overbrace{k_{deg}f_R[ObRa_i]}^{\text{Degradation}} \tag{7.11}$$

These unknown rate constants represent steps in the receptor regulation process which could be influenced by external factors known to alter leptin transport rate.

7.2.4 Computational Model

I developed a computational model (named “leppRedict”) of leptin endocytosis which can be fit to experimental data to allow the isolation and measurement of the unknown rate constants related to receptor regulation. Since the purpose of this model is the interpretation of *in vitro* experimental data, it was designed to model single compartment endocytosis rather than transcytosis. This is likely to exhibit reduced endocytosis in the presence of the same factors which impair transcytosis while reducing the complexity and difficulty of the experiment. If an *in vitro* endocytotic experiment fails to reproduce a steady state change in endocytosis rate in response to known mediators of transcytosis, it will be necessary to expand this model to consider full transcytosis by adding equation (7.3) to the model.

Of the six differential equations developed herein to model leptin transport and receptor regulation *in vivo*, only four (equations 7.2, 7.9, 7.10, 7.11) are necessary to model endocytosis *in vitro*, plus one additional equation to account for the varying leptin concentration in solution:

$$\frac{d[Lep]}{dt} = \overbrace{k_{-1}[LepObRa]}^{\text{Dissociation}} - \overbrace{k_1[Lep][ObRa]}^{\text{Binding}}$$

This system of five differential equations is solved numerically in the R programming language, using the "deSolve" library[17, 22]. This allows for predicting the dependent variables of bound, unbound, and intracellular receptors and free, bound, and internalized leptin for any arbitrary time course, initial conditions, and rate constants. By comparing each data point in this prediction to measured experimental values a "cost function" represents the distance between the prediction and reality:

$$Cost = \sum [predicted\ data - experimental\ data]^2$$

The R library "subplex" is utilized to find a set of rate constants which represent a global minimum of the cost function, using Tom Rowan's subspace-searching simplex algorithm[12, 18]. leppRedict then outputs these predicted rate constants and plots the time-course solution of the system of differential equations with these predicted values (see figure 7.4 for an example fit to artificial data with random error).

This software can be applied simultaneously to any arbitrary combination of treatment vs non-treatment dynamic *in vitro* leptin endocytosis experiments, to predict which rate constant(s) are being modified by the treatment.

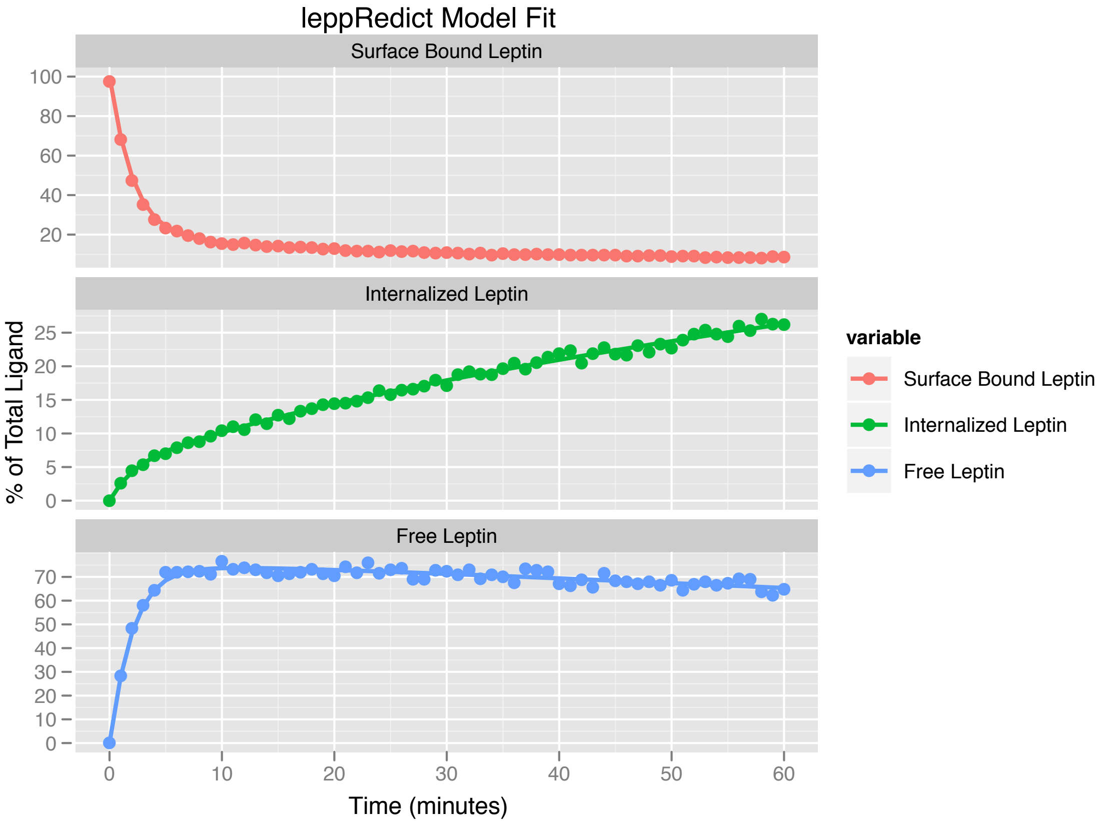


Figure 7.4: **An example fit of the leppRedict computational model to artificial data over a one hour time course beginning with all initial leptin bound to surface receptors.** Dots represent artificial data with random error applied, and solid lines show the model behavior as fit to these data. This figure illustrates the ability of the model to fit and visualize arbitrary data (and eventually experimental data); not a prediction of the system behavior.

7.3 Results and Discussion

7.3.1 Quasi-steady-state Behavior

Equation (7.6) predicts that this system will exhibit hyperbolic sensitivity of leptin transport rate ($d[Lep_t]/dt$) to blood leptin concentration ($[Lep]$) if the generalized mechanism of clarathrin-dependent receptor mediated transport (equation 7.1) from which it's derived accurately explains the underlying physiology. This prediction correlates closely with experimental data measuring unidirectional influx of radio-labeled $^{125}I - leptin$ across the blood-cerebrospinal fluid and blood-brain barriers using *in vivo* rat models[32]. Zlokovic et al. measured rapid but saturable uptake at physiological levels with a low v_{max} into the CNS at three sites (CSF, hypothalamus, and choroid plexus) and slow saturable uptake with a high v_{max} at above physiological levels in the hippocampus, cortex, and caudate nucelus (see figure 7.5 and table 7.1)[32]. This indicates significant differences in receptor expression and/or other factors involved in transport (such as abundance and internalization rate of clathrin-coated pits) at these sites.

According to the model presented herein, leptin levels must approach k_M for significant transport to occur. Human data suggests that plasma leptin levels in obese humans occur in the range of around $0 - 40 ng/ml$ (20.3 ± 5.8 for women, 5.8 ± 0.8 for men in ng/ml; $P = 0.001$)[20]. If the pattern of slow vs rapid transport is conserved between rats and humans, it is likely that significant transport for the purposes of lipostasis in living humans occurs only at sites of rapid transport (CSF,

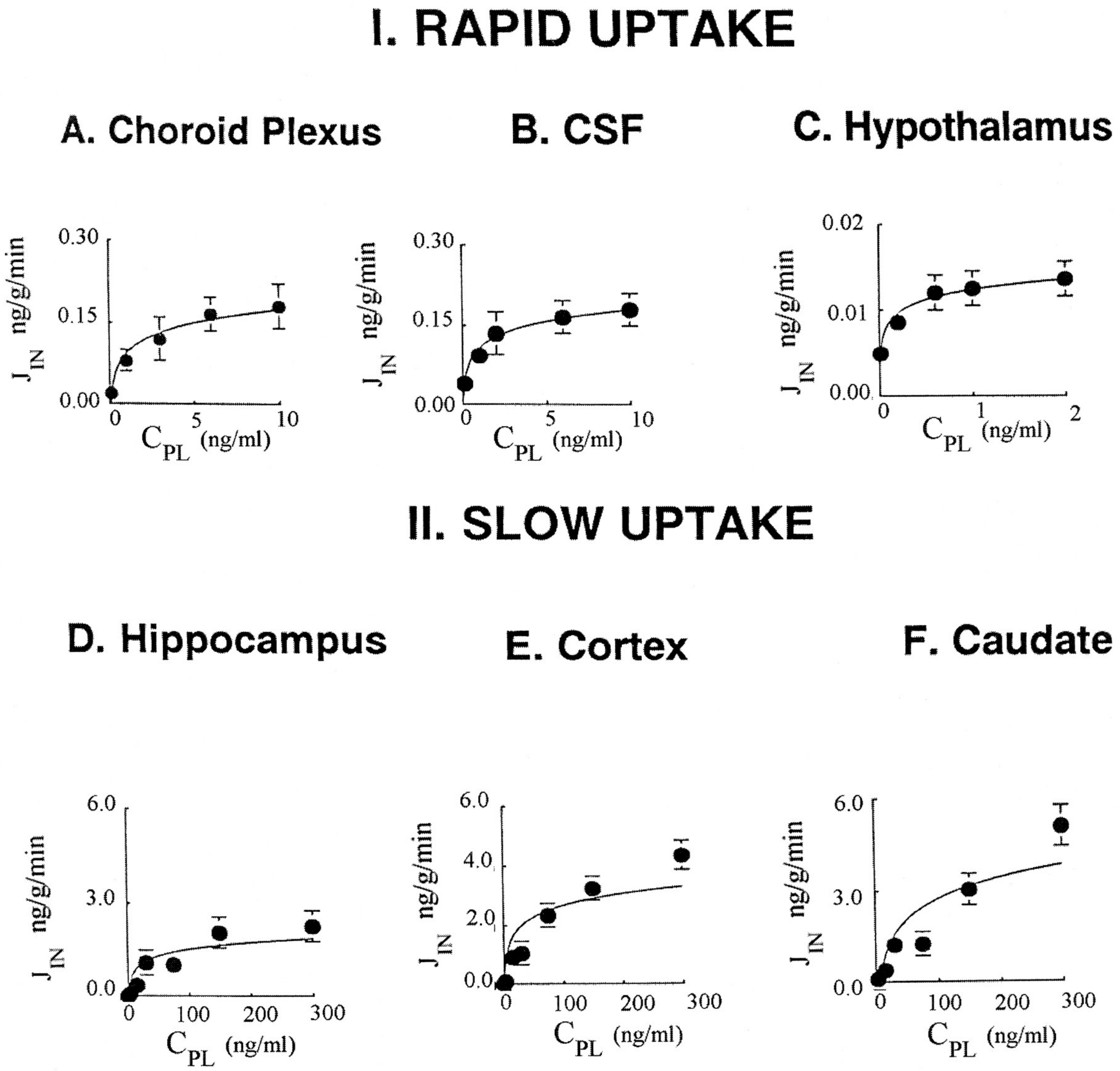


Figure 7.5: **Saturable response of transport rate to serum leptin levels at six sites from *in vivo* rat models.** Sites of slow uptake are unlikely to contribute significantly to leptin transport under physiological conditions. Figure republished from Zlokovic et al. 2000 with permission from the Endocrine Society[32].

Region	k_M (ng/ml)	v_{max} $(ng/g/min)$
CSF	$1.10 \pm 0.07^{\alpha}$	$0.20 \pm 0.007^{\alpha}$
Hypothalamus	0.23 ± 0.04	0.014 ± 0.003
Hippocampus	$88 \pm 10^{\alpha}$	$2.97 \pm 0.56^{\alpha}$
Cortex	$130 \pm 27^{\alpha}$	6.20 ± 0.50
Caudate Nucelus	345 ± 29	10.80 ± 2.60
Choroid Plexus	2.57 ± 0.48	0.23 ± 0.014

Table 7.1: **Saturable response of transport rate to serum leptin levels at six sites from *in vivo* rat models.** Choroid Plexus data represents binding to tissue rather than direct transport. Data from Zlokovic et al. 2000 [32].

Hypothalamus, Choroid Plexus), and therefore these sites are where deficient leptin transport pathophysiology is likely to occur.

7.3.2 Dynamic Behavior

The quasi-steady-state model of leptin transport presented herein predicts that transport rate could be mediated through mechanisms which affect either receptor-ligand affinity $K_D = k_{-1}/k_1$, rate of internalization (and therefore k_M), or regulation of short-form (transport) receptor expression $[ObRa]_0$ (and therefore v_{max}). Deviations from the hypothetical mechanism could also affect transport rate.

Applying conservation of mass to the hypothesized mechanism of ObRa receptor expression (production, internalization, recycling, and degradation) yields a system of differential equations which depend on unknown rate constants to describe the possible range of behavior within the context of this hypothesis. This system of simultaneous equations can be solved numerically with computer software (leppRedict) and fit to *in vitro* experimental data to estimate the values of these unknown rate constants.

While the underlying mechanism by which physiological factors (such as serum triglycerides, lipopolysaccharides, soluble leptin receptors, and α_1 adrenergics) mediate leptin transport are unknown, experimental data demonstrates a mathematical relationship between these factors and blood to cerebral spinal fluid (CSF) leptin concentration ratio[1, 2, 23]. Statistically significant relationships, such as a linear relationship between serum triglycerides and brain/serum leptin ratio ($r = -0.860, P < 0.05, n = 6$) have already been identified in literature but not yet synthesized into predictive models[2]. Since these data represent changes in the quasi-steady-state transport rate, they do not yield any insight into which step(s) in the biochemical mechanism of leptin transcytosis that they interact with, and therefore further experiments are necessary.

By comparing the dynamic behavior of the transport system in the presence and absence of the aforementioned factors known to influence leptin transport rate the altered rate constants can be identified and measured. This will provide new information about how these factors interact with the transport system.

7.4 Further Work

As described in the Background section, dynamic *in vitro* experiments were able to elucidate the underlying mechanism and rate constants for the epidermal growth factor (EGF) receptor-mediated endocytosis system[28]. I propose designing and conducting a similar series of experiments using existing laboratory cell lines which express the ObRa receptor (such as Madin-Darby canine kidney cells) to measure the underlying rate constants for both normal leptin transcytosis (as a control) and leptin transcytosis in the presence of a treatment compound known to mediate transport rate (such as free triglycerides)[9]. This could consist of either a two-compartment transcytosis experiment or a single-compartment endocytosis experiment depending on practical considerations. As with the EGF system, it will be necessary to stress the ObRa transport system under a variety of dynamic situations which have differing abilities to isolate specific rate constants in the reaction. For example, in the EGF system it was found that experiments that begin with cells incubated with ligand such that all initial ligand is bound to the cell surface are able to isolate and measure the dissociation rate constant, whereas experiments beginning with all ligand free in solution are more able to isolate and measure the binding rate constant. Both procedures were equally capable of measuring the rate of internalization[28]. After performing these experiments, a sensitivity analysis can be performed with existing software libraries (such as the R package "sensitivity")[16]. Sensitivity analysis systematically iterates the rate constants found by minimizing the cost function, to determine if the current solution is unique, or if the model could equally well be fit to the experimental data using a different set of rate constants. In the latter case, further experiments or refinements to the mathematical model would be necessary to isolate which rate constant(s) are being affected by the treatment protocol.

A better understanding of the underlying biochemical mechanism of known leptin transport mediators could aid in designing *in vitro* and *in vivo* experiments with mammalian tissue to confirm the underlying relationship, or to inform the development of metabolic syndrome therapies which influence these known mediators.

7.5 Bibliography

[1] W A Banks. Enhanced leptin transport across the blood-brain barrier by alpha 1-adrenergic agents. *Brain Res*, 899(1-2):209–217, Apr 2001.

[2] W A Banks, A B Coon, S M Robinson, A Moinuddin, J M Shultz, R Nakaoke, and J E Morley. Triglycerides induce leptin resistance at the blood-brain barrier. *Diabetes*, 53(5):1253–1260, May 2004.

[3] W A Banks, A J Kastin, W Huang, J B Jaspan, and L M Maness. Leptin enters the brain by a saturable system independent of insulin. *Peptides*, 17(2):305–311, 1996.

[4] V A Barr, K Lane, and S I Taylor. Subcellular localization and internalization of the four human leptin receptor isoforms. *J Biol Chem*, 274(30):21416–21424, Jul 1999.

[5] E. L. Cussler. *Diffusion: Mass Transfer in Fluid Systems (Cambridge Series in Chemical Engineering)*. Cambridge University Press, 1997.

[6] E O Diaz, A M Prentice, G R Goldberg, P R Murgatroyd, and W A Coward. Metabolic response to experimental overfeeding in lean and overweight healthy volunteers. *Am J Clin Nutr*, 56(4):641–655, Oct 1992.

[7] R C Frederich, B Löllmann, A Hamann, A Napolitano-Rosen, B B Kahn, B B Lowell, and J S Flier. Expression of ob mrna and its encoded protein in rodents. impact of nutrition and obesity. *J Clin Invest*, 96(3):1658–1663, Sep 1995.

[8] P. Hicks, C. Rolsten, D. Brizzee, and T. Samorajski. Age-related changes in rat brain capillaries. *Neurobiol. Aging*, 4:69–75, 1983.

[9] S M Hileman, J Tornoe, J S Flier, and C Bjorbaek. Transcellular transport of leptin by the short leptin receptor isoform obra in madin-darby canine kidney cells. *Endocrinology*, 141(6):1955–1961, Jun 2000.

[10] A J Kastin and W Pan. Intranasal leptin: blood-brain barrier bypass (bbbb) for obesity? *Endocrinology*, 147(5):2086–2087, May 2006.

[11] R F Keep and H C Jones. A morphometric study on the development of the lateral ventricle choroid plexus, choroid plexus capillaries and ventricular ependyma in the rat. *Brain Res Dev Brain Res*, 56(1):47–53, Oct 1990.

[12] Aaron A. King. *subplex: Subplex optimization algorithm*, 2008. R package version 1.1-3.

[13] N A King, M Hopkins, P Caudwell, R J Stubbs, and J E Blundell. Individual variability following 12 weeks of supervised exercise: identification and characterization of compensation for exercise-induced weight loss. *Int J Obes (Lond)*, 32(1):177–184, Jan 2008.

[14] Y L Lai, P M Smith, W J Lamm, and J Hildebrandt. Sampling and analysis of cerebrospinal fluid for chronic studies in awake rats. *J Appl Physiol*, 54(6):1754–1757, Jun 1983.

[15] M G Myers. Leptin receptor signaling and the regulation of mammalian physiology. *Recent Prog Horm Res*, 59:287–304, 2004.

[16] Gilles Pujol. *sensitivity: Sensitivity Analysis*, 2008. R package version 1.4-0.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[18] Thomas Harvey Rowan. *Functional stability analysis of numerical algorithms*. PhD thesis, Austin, TX, USA, 1990. UMI Order No. GAX90-31702.

[19] E Schiffer, E Van Gessel, and Z Gamulin. Influence of sex on cerebrospinal fluid density in adults. *Br J Anaesth*, 83(6):943–944, Dec 1999.

[20] M W Schwartz, E Peskind, M Raskind, E J Boyko, and D Porte. Cerebrospinal fluid leptin levels: relationship to plasma levels and to adiposity in humans. *Nat Med*, 2(5):589–593, May 1996.

[21] M W Smith and M Gumbleton. Endocytosis at the blood-brain barrier: from basic understanding to drug delivery strategies. *J Drug Target*, 14(4):191–214, May 2006.

[22] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving differential equations in r: Package desolve. *Journal of Statistical Software*, 33(9):1–25, 2010.

[23] H Tu, A J Kastin, H Hsuchou, and W Pan. Soluble receptor inhibits leptin transport. *J Cell Physiol*, 214(2):301–305, Feb 2008.

[24] S Uotani, C Bjorbaek, J Tornoe, and J S Flier. Functional properties of leptin receptor isoforms: internalization and degradation of leptin and ligand-induced receptor downregulation. *Diabetes*, 48(2):279–286, Feb 1999.

[25] M Van Heek, D S Compton, C F France, R P Tedesco, A B Fawzi, M P Graziano, E J Sybertz, C D Strader, and H R Davis. Diet-induced obese mice develop peripheral, but not central, resistance to leptin. *J Clin Invest*, 99(3):385–390, Feb 1997.

[26] T A Wadden. Treatment of obesity by moderate and severe caloric restriction. results of clinical research trials. *Ann Intern Med*, 119(7 Pt 2):688–693, Oct 1993.

[27] M Y Wang, Y T Zhou, C B Newgard, and R H Unger. A novel leptin receptor isoform in rat. *FEBS Lett*, 392(2):87–90, Aug 1996.

[28] C. M. Waters, K. C. Oberg, G. Carpenter, and K. A. Overholser. Rate constants for binding, dissociation, and internalization of EGF: effect of receptor occupancy and ligand concentration. *Biochemistry*, 29:3563–3569, Apr 1990.

[29] H S Wiley, S Y Shvartsman, and D A Lauffenburger. Computational modeling of the egf-receptor system: a paradigm for systems biology. *Trends Cell Biol*, 13(1):43–50, Jan 2003.

[30] B E Wisse and M W Schwartz. Does hypothalamic inflammation cause obesity? *Cell Metab*, 10(4):241–242, Oct 2009.

[31] Y Zhang, R Proenca, M Maffei, M Barone, L Leopold, and J M Friedman. Positional cloning of the mouse obese gene and its human homologue. *Nature*, 372(6505):425–432, Dec 1994.

[32] B V Zlokovic, S Jovanovic, W Miao, S Samara, S Verma, and C L Farrell. Differential regulation of leptin transport by the choroid plexus and blood-brain barrier and high affinity transport systems for entry into hypothalamus and across the blood-cerebrospinal fluid barrier. *Endocrinology*, 141(4):1434–1441, Apr 2000.

Chapter 8

Conclusion and Future Vision

8.1 Overview

At first glance, the conceptual continuity between the chapters of this dissertation may not be readily apparent, but they can be seen as different aspects of a single long-term goal, of integrating relevant biological data from diverse sources together in a machine readable manner, and leveraging it towards the problem of identifying safe and effective small molecule therapies. This goal entails leveraging small molecule structures, bioactivity data, and complex systems data about the target organism together in a meaningful way. The chapter on ChemMine Tools represents an effort to make small molecule structural analysis tools widely available, in a manner that can be used to identify chemical features with therapeutic relevance. These tools provide both a web interface, and a programmable R language interface, to make them available to a diverse user base. The chapters on bioassayR represent a complementary effort to enable large scale cross-target analysis of small molecule High Throughput Screening (HTS) results from large data resources such as PubChem BioAssay[28]. The chapter on target selectivity demonstrates how bioassayR can be leveraged to systematically analyze these activity data, and identify patterns of drug-likeness that can inform drug discovery efforts. Lastly, the chapter on leptin biotransport demonstrates a method to integrate experimental knowledge into a coherent mathematical model of a potential target pathway, that captures the underlying dynamics of the system. An accurate dynamic model of a disease related pathway presents an opportunity to identify druggable targets *in silico*, and rank candidates identified from HTS data by there predicted efficacy against the dynamic system. Each chapter in this dissertation contains a detailed discussion of it's findings, but here I review them again in the context of there contributions towards the long-term goal of a unified drug discovery workflow.

8.2 ChemMine Tools

While most of the work in this dissertation is focused on experimental bioactivity profiles, structural similarity based small molecule hit expansion by methods such as those provided by ChemMine Tools has shown complementary effectiveness with bioactivity based hit expansion, producing a largely different yet accurate set of hit predictions[25, 22, 8, 5]. Additionally, *in silico* structural analysis has a much lower barrier to entry, as it doesn't require large quantities of high throughput screening data, and can therefore be practically performed earlier in a drug discovery pipeline.

ChemMine Tools is a modular web framework for analyzing small molecule structures. The tools all share a simple workflow that takes a set of small molecule structures and user options as input, and produces an output file with the results of this analysis. As most cheminformatics software tools can fit into a workflow of this type, adding new tools trivially involves writing a short YAML formatted text file that describes the input and output formats of each tool. From this data, the ChemMine Tools framework automatically generates an appropriate job submission form, and links the output to a web viewer appropriate for each output format, to visualize the results. For R language users, an example function call is automatically created for launching the job, and the results are translated back into a standard R object. Consequently, this substantially lowers the previously labor intensive barrier to entry in running complex cheminformatics pipelines, without the need for end users to individually develop converters and report generators for each step. It also allows users free access to the powerful compute resources that host ChemMine Tools.

The functionalities provided by ChemMine Tools integrate into a drug discovery pipeline in many ways. For example, the Compound Workbench allows users to create, organize, and explore an *in silico* compound library, while adding or removing small molecules based on the results of other tools. The Clustering Workbench provides many different ways to cluster and organize compounds by similarity, including the ability to view structural clusterings alongside heatmaps of user uploaded activity data, and visually identify patterns. The Properties Toolbox provides access to a wide array of physicochemical property calculators, and can be used to correlate these properties with experimental results, or to filter a compound library based on desired properties such as the Lipinksi Rule of Five[17]. The Similarity Workbench provides pairwise similarity comparisons that can be both visually and quantitatively explored to, for example, identify the common shared substructure between a set of small molecules that are known to both induce a desired phenotype. Lastly, the Search Toolbox provides two powerful algorithms for rapidly identifying small molecules similar to a desired query. These tools search the entire PubChem compound structure database, and provide an automated import feature, where the small molecule results of a search can be retrieved from the PubChem servers remotely, and added to a users Compound Workbench[15].

8.3 bioassayR

bioassayR is an open source R language software package distributed through the Bioconductor project, which analyzes data from a large number of screening experiments to identify target selective drug candidates, druggable protein targets, and patterns of bioactivity that lead to drug-likeness[9]. By simultaneously analyzing data from both users custom small molecule assays and public databases, bioassayR can identify protein targets accessible to small molecule perturbation, elucidate novel mechanisms of action for bioactive molecules, and predict undesirable off-target effects.

bioassayR provides a large number of functions that can contribute to drug discovery workflows by informing them with large bioactivity data. For example, given a protein target or pathway of interest, bioassayR will provide a list of candidate small molecules active and/or selective against these targets. If no screening data exists for the targets of interest, bioassayR can identify small molecules active against targets which share common protein domains and/or sequence similarity to those of interest. The fingerprint features in bioassayR allow users to cluster compounds by bioactivity patterns, and also to perform bioactivity based hit expansion, which identifies compounds sharing cross-target activity patterns with a set of query compounds that are known to have desirable bioactivities. bioassayR also provides a Bayesian statistical model of target selectivity, which can be used to rank, and filter compounds by selectivity or it's inverse (promiscuity) in a manner that limits the bias induced by compounds screened against differing quantities of different targets.

The "results of a high-content screen compared to single target bioassays of the same pathway" table in the bioassayR Supporting Information chapter provide strong validation of the capabilities bioassayR provides to predict *in vivo* bioactive molecules from large *in vitro* bioactivity data. Out of a library of 61605 compounds, bioassayR predicts 1749 actives likely to exhibit activity in an assay which uses a luciferase reporter to measure NF-kappaB activation in HEK-293 cells. Of these, 73 were experimentally active showing a prediction rate of 4.12%, compared to only 0.21% in the entire library of 61605 compounds. Additionally, bioassayR provides candidate protein targets for the *in vivo* active compounds.

8.4 Large-scale Bioactivity Analysis of the Druggable Proteome

The systematic bioactivity analysis of "highly screened" small molecules in PubChem BioAssay identifies many surprising patterns of bioactivity for both existing FDA approved drugs, and other compounds which can potentially inform further drug discovery efforts. For example, there is significant debate in the existing literature about how the concept of polypharmacology should be applied to drug discovery[12, 3, 10, 20, 30, 16, 18, 1, 11]. Given that existing FDA approved drugs tend to show activity against a large number of targets on average, does this mean that these drugs mostly act by polypharmacological mechanisms? If so, this raises the question of which drug discovery strategy to utilize going forward, *i.e.* should we look for highly selective molecules which interact with a single therapeutic target, or look for polypharmacological compounds which interact with many nodes in a complex biological network?

The results of this analysis provide several interesting observations, including new information that may help lead to a more nuanced view of these polypharmacological questions. First, while FDA approved drugs tend to exhibit activity a median of 7 targets, vs 2 for other bioactive compounds, this number decreases substantially to 4 targets when targets sharing common protein domain families are merged, while the median number of actives for other compounds remains at 2. Additionally, when the bioactivities of the FDA approved drugs are stored in a compound-target binary matrix, a biclustering analysis yields many dense biclusters, where a common set of drugs are active against a common set of targets which share common domains and molecular function annotations. This shows that many of the FDA approved drugs with a large number of active targets are not selective at the individual protein target level, but exhibit selectivity for a class of related proteins. While this provides a partial explanation for the observed multi-target activity of FDA approved drugs, further analysis is necessary to determine to what extent some of these compounds may also be targeting multiple nodes in a target system, via a pharmacological mechanisms of drug action. This analysis may become more practical in the future, as the dynamics and connections of these target systems are better understood.

Additionally, the FDA approved drugs exhibit a wide range of target selectivities due to many different mechanisms of activity, which is highlighted by the analysis which individually looks at the selectivity of compounds active against distinct protein domains. While proteins with some domains tend to have highly selective active small molecules, others tend to have highly promiscuous active compounds. Individually looking at drugs by class shows many diverging trends of target selectivity based on the therapeudic strategy of each drug. For example, large natural antibiotics such as streptomycin exhibit high selectivity for bacterial ribosomal RNA and tend to show no or very limited activity against protein targets. Conversely, many chemotherapy drugs such as broad spectrum kinase inhibitors show activity against nearly all tested targets. As such, taking the average or mean target selectivity of drugs with such widely different strategies of activity may provide limited insights into how to treat a particular disease.

This chapter includes a Bayesian probablistic model of cross-reactivity, which yields a probability distribution over the expected hit ratios for a large number of compounds. The hit ratio is the expected fraction of actives, if a compound were screened against an entire proteome, or large number of active targets. Using this model, we quantify and visualize the fact that the FDA approved drugs exhibit a much higher hit ratio than other compounds, demonstrating that they are enriched for activity against many targets. We also compare the hit ratio probabilities for two common methods of determining promiscuous compounds, which often interfere with high throughput screening efforts by exhibiting non-selective activity. We find that Pan Assay Interference Compounds (PAINS) accurately identify compounds with a much higher average hit ratio, while promiscuous aggregators had only a sightly higher hit ratio overall, but showed greater ability to identify the most promiscuous compounds[2, 7, 19]. We also build a Target Protein (TP) network, where all protein targets are represented as nodes, as are connected if they share a high percentage of active compounds. By leveraging the probabilistic cross-reactivity model, it is possible to remove a small number of the most promiscuous compounds which results in the removal of a comparatively large number of likely spurious nodes in the network. This suggests that the cross-reactivity model has the potential to improve the quality and utility of bioactivity profile comparisons, by removing activity outcomes from unselective compounds.

We also find that the distribution of target selectivities (total active targets) for individual small molecules follows a stretched exponential distribution as commonly observed in many natural multiplicative processes, suggesting the hypothesis that active drug-target pairs may result from a large number of different thermodynamic barriers, each with different probabilities. While many of these barriers may involve already well understood and studied mechanisms, it also opens the possibility of computationally screening for computable physicochemical properties which may identify novel barriers, with potential utility for finding additional predictors of bioactivity which are computable *in silico*.

8.5 Leptin Transcytosis

This chapter provides an example of extracting a detailed qualitative explanation of a biochemical mechanism from existing literature, and translating this system into a machine readable system of differential equations. I use apply information about the physical structure and chemical kinetics of this system to make simplifications and idealizations that render the model computationally tractable. For example, by estimating the The Damköhler number I show that the rate of transport is not limited by diffusion, and therefore diffusion terms can be eliminated from the model without negatively affecting the resulting predictions.

The model predicts saturable transport consistent with the Michaelis-Menten equation of enzyme kinetics, which agrees with experimental data collected from model organisms. This quantitative model provides two key opportunities to further study the process of leptin resistance, which may lead to obesity. First, it can inform the design of simple biochemical experiments that, when fit to the model, will estimate currently unknown parameters in this system, such as reaction rate constants.

Secondly, it can make predictions regarding the downstream effects of system perturbation such as a small molecule drug which inhibits a specific enzyme in the system. This allows for the possibility of *in silico* small molecule drug screening informed by the dynamics of the system.

8.6 Future Vision

The best material model of a cat is another, or preferably the same, cat.

Norbert Wiener and Arturo Rosenblueth[27]

Natural discovery is fundamentally a process of modeling, if only in the narrow technical sense that we perceive natural phenomenon, construct a mental image of how the process works, and then leverage that understanding to perform experiments and observations that further update our understanding. As our understanding of biology becomes more complex and quantitative, it will become increasingly necessary to formalize our understanding in mathematical and computer models that can translate our current state of knowledge into predictions that are not intuitively obvious. As George E. P. Box said, “all models are wrong but some are useful”[4]. When working with a mental model of a natural process, we can be careful to keep in mind the limitations of our model which result from limited understanding and evidence. This presents a serious challenge when translating this understanding into a concrete, deterministic set of formulas as in the Leptin Transport chapter of this dissertation. To run this type of model, we must choose one specific discrete explanation and set of parameters for the underlying system. This results in a model that doesn't truly reflect our understanding of the system, as it discards one of the most important categories of knowledge- where the gaps in our understanding lie. Bayesian probability theory provides a unique approach to this problem as it rigorously extends discrete logic into the realm of continuous probability[13]. This allows us to build models which incorporate uncertainty about the underlying system. If our knowledge about a biological system is encoded as a probability distribution over different possible models and model parameters, the outcome of the computer model is a probability distribution over expected system behaviors that can be experimentally verified, and used to further improve our knowledge[24]. This has been understood since the time of Pierre-Simon Laplace, but recent advances in computing power have extended it's potential utility to a much larger set of problems. This framework provides several key opportunities: (i) we can systematically identify which currently missing information would be most valuable at increasing our ability to make meaningful predictions; (ii) we can simultaneously inform our model with experimental data from a wide range of sources, represented in a standard way as probability distributions; (iii) we can use Monte Carlo algorithms to estimate the posterior distribution of complex models; (iv) we can model phenomena to which the concept of frequency is not applicable.

Many advances have been recently made towards increasingly accurate computer simulation of biomolecular feedback systems, stochastic biochemical kinetics, and metabolic fluxes[26, 6, 23]. These modeling approaches have also been integrated into multi-level whole-cell microbial models which make accurate phenotypic predictions[14]. Software systems such as TherapySim have also demonstrated the potential of dynamic target system models to make novel therapeutic target and small molecule drug predictions[21]. As shown in the Bayesian compound promiscuity probability model of the cross-target analysis chapter, I have endeavored to curate the available bioactivity data in a machine readable manner, as probability distributions reflecting our uncertainty due to missing data and experimental error. This includes both the probability that a given compound-target active score is a "true positive" as well as a continuous probability distribution over the hit ratio θ, representing the expected fraction of actives among the untested protein targets for each compound. These probabilities can be further improved by adding additional complexity and nuance to the underlying model, *e.g.* treating protein target families independently, while considering the level of correlations between them, as well as estimating sensitivity and specificity separately and individually for each assay or assay type. These distributions can then be sampled in combination with models of target systems and living cells, providing a posterior probability distribution over the space of inducible phenotypes. These probabilistic models, along with the uncertainty present in the underlying bioactivity data will allow us to rank and quantify drug candidates based on the probability of inducing a desired phenotype, as well as identify which new experiments are likely to be most informative. In an appropriate context, it may also become possible to discriminate between inert, adverse, and desirable patterns of polypharmacology, in the context of a molecules predicted effects on a biological control system. I hope that this will enable the discovery of safer and more effective drug candidate predictions, which incorporate an increasingly accurate understanding of the dynamics that emerge from complex biological systems. Despite a large body of existing work, developing this type of probabilistic *in silico* drug screening will be extremely challenging and will require a deeper understanding of cell biology, as well as a large number of new computational and mathematical tools[29]. Nevertheless, it is a bright future with manifold implications for improving human health, that I am excited to work towards.

8.7 Bibliography

[1] Ravi Iyengar Aislyn DW Boran. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 2010.

[2] Jonathan B Baell and Georgina A Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.*, 2010.

[3] Jürgen Bajorath. Analyzing Promiscuity at the Level of Active Compounds and Targets. *Molecular Informatics*, 2016.

[4] George E P Box. *Robustness in the Strategy of Scientific Model Building.* Academic Press, 1979.

[5] Xin Chen and Charles H Reynolds. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.*, 2002.

[6] Domitilla Del Vecchio and Richard M Murray. *Biomolecular Feedback Systems.* Princeton University Press, 2014.

[7] Brian Y Feng, Anton Simeonov, Ajit Jadhav, Kerim Babaoglu, James Inglese, Brian K Shoichet, and Christopher P Austin. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.*, 2007.

[8] Peter Gedeck, Bernhard Rohde, and Christian Bartels. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.*, 2006.

[9] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 2004.

[10] Ye Hu and Jürgen Bajorath. Compound promiscuity: what can we learn from current data? *Drug Discov. Today*, 2013.

[11] Ye Hu and Jürgen Bajorath. What is the Likelihood of an Active Compound to Be Promiscuous? Systematic Assessment of Compound Promiscuity on the Basis of PubChem Confirmatory Bioassay Data. *AAPS J*, 2013.

[12] Swarit Jasial, Ye Hu, and Jürgen Bajorath. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS ONE*, 2016.

[13] E T Jaynes and G Larry Bretthorst. *Probability Theory.* Cambridge University Press, 2003.

[14] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 2012.

[15] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem Substance and Compound databases. *Nucleic Acids Res.*, 2016.

[16] Joseph Lehár, Andrew S Krueger, William Avery, Adrian M Heilbut, Lisa M Johansen, E Roydon Price, Richard J Rickles, Glenn F Short III, Jane E Staunton, Xiaowei Jin, Margaret S Lee, Grant R Zimmermann, and Alexis A Borisy. Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Rev. Drug Discovery*, 2009.

[17] C A Lipinski, F Lombardo, B W Dominy, and P J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 2001.

[18] Jin-Jian Lu, Wei Pan, Yuan-Jia Hu, and Yi-Tao Wang. Multi-Target Drugs: The Trend of Drug Research and Development. *PLoS ONE*, 2012.

[19] Susan L McGovern, Brian T Helfand, Brian Feng, and Brian K Shoichet. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.*, 2003.

[20] A A Moya-Garcia and J A G Ranea. Insights into polypharmacology from drug-domain associations. *Bioinformatics*, 2013.

[21] Daniel Nicklas and Leonor Saiz. In silico identification of potential therapeutic targets in the TGF-β signal transduction pathway. *Mol. BioSyst.*, 2014.

[22] Tudor I Oprea. Chemical space navigation in lead discovery. *Current Opinion in Chemical Biology*, 2002.

[23] Jeffrey D Orth, Ines Thiele, and Bernhard O Palsson. What is flux balance analysis? *Nat. Biotechnol.*, 2010.

[24] V Periwal. Bayesian inference of biological systems: the logic of biology. *System modeling in cellular biology: from concepts to ...*, 2010.

[25] Paula M Petrone, Benjamin Simms, Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhan Deng, John W Davies, Jeremy L Jenkins, and Meir Glick. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.*, 2012.

[26] Christopher V Rao and Adam P Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.*, 2003.

[27] A Rosenblueth and N Wiener. The Role of Models in Science. *Philosophy of science*, 1945.

[28] Yanli Wang, Tugba Suzek, Jian Zhang, Jiyao Wang, Siqian He, Tiejun Cheng, Benjamin A Shoemaker, Asta Gindulyte, and Stephen H Bryant. PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, 2013.

[29] Darren J Wilkinson. *Stochastic Modelling for Systems Biology, Second Edition.* CRC Press, 2011.

[30] Wendy Wolfson. Speed Dating for Molecules CombinatoRx Looks for That Special Synergy. *Chem. Biol. (Oxford, U. K.)*, 2006.