

THE EFFECT OF DATA CURATION ON THE ACCURACY OF QUANTITATIVE  
STRUCTURE-ACTIVITY RELATIONSHIP MODELS

Andrew Dale Fant

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in  
partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of  
Pharmacy (Chemical Biology and Medicinal Chemistry)

Chapel Hill  
2015

Approved by:

Alexander Tropsha

Timothy C. Elston

Andrew L. Lee

Ivan I. Rusyn

Scott F. Singleton

© 2015  
Andrew Dale Fant  
ALL RIGHTS RESERVED

## ABSTRACT

Andrew Dale Fant: The Effect of Data Curation on the Accuracy of Quantitative Structure-Activity Relationship Models  
(Under the direction of Alexander Tropsha)

In the 33 years since the first public release of GenBank, and the 15 years since the publication of the first pilot assembly of the human genome, drug discovery has been awash in a tsunami of data. But it has only been within the past decade that medicinal chemists and chemical biologists have had access to the same sorts of large-scale, public-access databases as bioinformaticians and molecular biologists have had for so long. The release of this data has sparked a renewed interest in computational methods for rational drug design, but questions have arisen recently about the accuracy and quality of this data. The same question has arisen in other scientific disciplines, but it has a particular urgency to practitioners of Quantitative Structure-Activity Relationship (QSAR) modeling. By its nature QSAR modeling depends on both activity data and chemical structures. While activities are usually expressed as numerical scalar values, a form ubiquitous throughout the sciences, chemical structures (especially that must be interpretable as such by computer software) are stored in a variety of specialized formats which are much less common and mostly ignored outside of cheminformatics and related fields.

While previous research has determined that a 5% error rate in data being used for modeling can cause a QSAR model to be non-predictive and useless for its intended purpose, and workflows have been proposed which reduce the effect of inconsistent chemical structure representations on model accuracy, a fundamental question remains: “how

accurate are the structure and activity data freely available to researchers?” To this end, we have undertaken two surveys of data quality, one focusing on chemical structure information in Internet resources and a second examining the uncertainty associated with compounds reported in the medicinal chemistry literature as abstracted in ChEMBL. The results of these studies have informed the creation of an improved workflow for the curation of structure-activity data which is intended to identify problematic data points in raw data extracted from databases so that an expert human curator can examine the underlying literature and resolve discrepancies between reported values. This workflow was in turn applied to the creation of two QSAR models that were used to implement a virtual screen seeking molecules capable of binding to both the serotonergic reuptake transporter and the alpha 2a receptor. While no suitable compounds were identified in the initial screening process, regions of chemical space that may yield truly novel alpha 2a receptor ligands have been identified. These regions can be targeted in future efforts.

Basing data curation workflows on manual processes by human curators is not particularly viable, as humans have a tendency to introduce errors by inattention even as they identify and repair other problems. Computers cannot effectively curate data either. While they are highly accurate when programmed properly, they lack human creativity and insight that would allow them to determine which data points represent truly inaccurate information. In order to effectively curate data, humans and computers must both be incorporated into a workflow that harnesses their strengths and limits their liabilities.

For Phyllis  
Who Understood Why

## ACKNOWLEDGEMENTS

It can be a daunting experience to look back from the near-completion of a doctoral dissertation and take stock of all the people who have helped me to reach this point. I am grateful beyond words to everyone who has helped, and I apologize to anyone who may have inadvertently been omitted.

Before I ever set foot in a university classroom, I was fortunate to have known some excellent science teachers whose collective wisdom I still draw upon: Sherry Grimm, John Carpenter, Martha Swacen, and Bruce Secker. They helped set me on this path.

As I went on to study chemistry, several faculty served as my mentors and research advisors. Paul Endres, David Newman, Joel Liebman, and Tom Kinstle: thank you for the time and effort expended in teaching me to think like a chemist.

During my first career, I met many molecular modelers and cheminformaticians who have been role models and set a high standard for what I hope to be professionally: Pat Walters, Mark Murcko, Rajarshi Guha, Ken Mattes, and Joe Leonard have all generously shared their advice over the years. I am the richer for it and appreciate it greatly.

Steven Peseckis asked me why I *wasn't* in graduate school when I approached him about a collaboration after I moved to Toledo in 2007. That comment set all of this in motion. Thank you, Steve. Similarly, the Medicinal and Biological Chemistry department at Toledo took it in relatively good humor when they found out that a self-funded student

would be taking courses with them, and the Department of Neurosciences welcomed me as an unexpected neighbor for that academic year.

As an intern in structural biology at Boehringer Ingelheim during the summer of 2012, Jörg Bentzein, Ingo Mügge, Prasenjit Mukherjee, and Matthias Zentgraf welcomed me into the modeling group and helped stretch my technical skills in new directions. Danke.

The work presented in chapter two was performed in cooperation with several collaborators beyond the MML at UNC. Christopher Southan (AstraZeneca), Andrey Erin (ACD/Labs), Tony Williams and David Sharpe (RSC), Jordi Mestres (IMIM) and Ricard Garcia (Chemotargets, S.L) all contributed their own preferred methods for chemical structure searching and their results for the specific drug list examined. Phyllis Pugh provided workflow graphics and statistical consulting. OpenEye Scientific Software, ChemAxon, and ACD/Labs donated access to their software suites that were used for various parts of the analysis workflow.

Over the past couple of years, I have travelled to Birmingham on a semi-regular basis. The members of the Sarah Clinton and Ilan Kerman lab in Psychiatry at UAB have been gracious hosts during that time, welcoming me to set up shop in unused corners and use their Internet Access. They have also allowed me to become involved in some of their research and exposed me to a very different set of problems from those I have worked on in this work.

Because I have participated in the certificate program in Bioinformatics and Computational Biology, I have been fortunate enough to feel like I have a home in two different departments. Through the program, I have made many, many friends. In particular, thank you to all the regulars at the BCB First Monday social hour for all the good memories.

Being a graduate student means that there is a dissertation committee casting a long shadow over one's research. Thank you to Drew Lee, Tim Elston, Scott Singleton and Ivan Rusyn for agreeing to serve on mine and all their intense questioning.

Earning a PhD in the sciences is more of a team sport than a solitary pursuit. I have been fortunate to have laughed, drank, and even occasionally worked with some of the most demented labmates ever known, and I am the richer for it. Particular thanks go to Chris Grulke, Denis Fourches, Steve Bush, Eugene Muratov, Denise Polhaus, Simon Wang, Nancy Baker, and Mary La. Special thanks are due to Alecks Sedykh for alerting me to an open research fellowship that I now hold.

During my time in Chapel Hill, Mike Jarstfer has been a mentor, and more importantly a friend. Whether discussing applied pedagogy, enzyme kinetics, or the art of building a career in the sciences, he has shared pieces of himself that will stay with me for a long time.

Alex Tropsha took a chance on accepting a very non-traditional graduate student into the program and his lab and not only indulged my scattershot curiosity but financially supported it. I owe him many thanks.

When I announced that I was going back to school, my parents, Molly and Fred Fant, and my sister Emily were never anything less than enthusiastic. Since then, they have continued to express their encouragement and support (moral, intellectual, and occasionally financial). Without them, this process would have been much harder; it might not have even have happened at all.



Finally, above anyone else, I have to express my deepest love and gratitude to my wife, Phyllis. When we married, neither of us suspected that “for better or for worse” would involve us living separated by 700 miles for most of 6 years, but she has seen me through it with more support, understanding, and affection than I had any right to expect. She, more than anyone else, has made this work possible.

## TABLE OF CONTENTS

LIST OF TABLES .....	xiv
LIST OF FIGURES .....	xvi
LIST OF ABBREVIATIONS AND SYMBOLS .....	xviii
CHAPTER 1: INTRODUCTION .....	1
References .....	14
CHAPTER 2: CHEMICAL STRUCTURE ACCURACY .....	16
Summary .....	16
Introduction .....	18
Methods.....	30
Creation of initial drug name list.....	30
University of North Carolina Workflow .....	31
Royal Society of Chemistry Workflow .....	34
AstraZeneca Workflow.....	36
IMIM Workflow .....	37
Structural Comparisons.....	38
Statistical Methods .....	38
Results .....	39
Comparison of the ability of workflows to generate correct structures .....	38
Evaluation of chemical structure repositories .....	39
Discussion .....	42

Categorizing errors .....	42
Comparisons between workflows .....	47
Standardizing electronic representations .....	49
Conclusions .....	51
References .....	56
CHAPTER 3: BIOLOGICAL DATA CURATION .....	59
Summary .....	61
Introduction .....	61
Methods.....	67
ChEMBL Logical Organization .....	67
Initial Data Assembly.....	72
Refinement of Data Assembly .....	73
Results .....	76
Distribution of magnitude of differences in paired binding affinity measurements.....	76
Distribution of binding affinity data .....	77
Error types extracted from primary examples .....	84
Error estimates between biological replicates .....	85
Conclusions .....	88
Implications for curation of SAR data/Heuristics for curation.....	91
Proposed Workflow .....	92
References .....	109
CHAPTER 4: QSAR MODELS OF $\alpha$ 2a ADRENERGIC RECEPTOR AND SEROTONIN REUPTAKE TRANSPORTER BINDING.....	111
Summary .....	111
Introduction .....	114

Methods.....	119
Data set extraction and curation .....	119
QSAR workflow.....	121
Data extraction in R and Caret .....	121
Descriptor generation .....	122
Data set splitting .....	122
k-Nearest Neighbors (kNN) modeling .....	123
Random Forest (rf) modeling.....	123
Support Vector Machine (SVM) classifiers .....	124
Model assessment .....	124
Genetic Algorithm variable selection .....	125
Prediction of activity.....	126
Virtual screening.....	127
Results .....	127
$\alpha$ 2a.....	128
Extraction and curation of data.....	128
Model optimization for $\alpha$ 2a .....	129
Production models.....	129
Y randomization.....	132
Regression modeling .....	132
Semi-curated data set .....	133
Virtual screening for $\alpha$ 2a .....	133
SERT.....	135
Data extraction and curation .....	135
Classifier modeling .....	136

Y randomization.....	136
Regression modeling .....	132
Virtual screening for SERT.....	138
Overlap between $\alpha$ 2a adrenergic hits and SERT hits .....	139
Consensus model performance on training data .....	140
Discussion .....	142
Deduplication strategy and relevance .....	143
Conclusions .....	146
References .....	148
CHAPTER 5: CONCLUSIONS .....	152
Towards rational data curation of biochemical affinity data .....	152
Data density and the inclusion of data from multiple sources in a single modeling set .....	154
Further Work and Directions .....	156
References .....	160
APPENDIX 1: NAMES OF DRUGS FOR INTERNET STRUCTURE SEARCH .....	161
APPENDIX 2: GOLD LIST FINAL CONSENSUS DRUG STRUCTURES .....	165
APPENDIX 3: SAMPLE SQL SOURCE FOR CHEMBL DATA EXTRACTION .....	184
APPENDIX 4: SOURCE CODE FOR ATOM-PAIR DESCRIPTOR CALCULATION .....	188
APPENDIX 5: SOURCE CODE FOR VARIABLE SELECTION QSAR MODEL CONSTRUCTION .....	191

## LIST OF TABLES

Table 1.1:	Raw occurrence of different error types and specific error rate estimates for each data source considered .....	11
Table 2.1:	Incorrect and correct structures for neomycin .....	24
Table 2.2:	Correct and incorrect structures of bosuntinib .....	26
Table 2.3:	Compounds for which ChemDraw was unable to generate correct systematic names.....	34
Table 2.4:	Summary of search results by each team compared to the consensus gold list .....	39
Table 2.5:	Number of hits returned for compounds on consensus gold list for different open access structural databases, and number of correct hits returned .....	40
Table 2.6:	Sample incorrect and correct structures for several compounds retrieved as part of the initial structure resolution process.....	42
Table 2.7:	Sample incorrect and correct structures found in public structure sources.....	38
Table 2.8:	Errors and corresponding correct structures introduced by manual intervention in the curation workflow .....	47
Table 2.9:	Relative frequency of numbers of groups reporting an incorrect structure for each search term .....	49
Table 3.1:	ChEMBL 14 Confidence Scores for target assignment quality assessment .....	69
Table 3.2:	ChEMBL Relationship Types.....	70
Table 3.3:	Number of targets, ligands, and documents in different subsets of ChEMBL 14 .....	78
Table 3.4:	Counts of errors found in a small subset of problematic entries in ChEMBL.....	84
Table 3.5:	Contingency for $\delta$ relative to $a$ and $b$ in hypothetical affinity pair .....	87
Table 3.6:	Simulated data for affinities of an antagonist binding to nicotinic acetylcholine receptor $\alpha 7$ in four different species.....	95

Table 3.7:	Simulated data for binding affinity for a hypothetical antagonist to various subtypes of serotonergic receptors.....	96
Table 3.8:	Simulated affinity data for multiple assays .....	98
Table 4.1:	Sizes of Data sets for Alpha-2a and SERT at multiple stages in the curation process.....	127
Table 4.2:	Summary of $\alpha$ 2a adrenergic models.....	131
Table 4.3:	Summary of serotonin reuptake transporter (SERT) models .....	137
Table 4.4:	Training compounds consistently mis-predicted in final consensus models for $\alpha$ 2a and SERT .....	141
Table 4.5:	Most commonly occurring descriptors in models.....	144

## LIST OF FIGURES

Figure 2.1:	Chemical structure resolution workflow as implemented at UNC .....	32
Figure 2.2:	Chemical structure resolution workflow as implemented by RSC .....	34
Figure 2.3:	Graphical comparison of performance of different sources against consensus gold list.....	41
Figure 2.4:	Summary comparison of accuracy of different structural curation workflows .....	53
Figure 2.5:	Chemical Structure of Taxol as displayed on ChemSpider and as recreated with SMILES string obtained from Chemspider via ChemDraw and MarvinSketch .....	54
Figure 3.1:	ERD for core tables of ChEMBL 14, with all columns and constraints shown .....	71
Figure 3.2:	Distribution of absolute activity differences for biological replicate pairs identified in ChEMBL.....	76
Figure 3.3:	Number of ligands tested against individual targets .....	81
Figure 3.4:	Number of times different compounds are referenced in distinct documents.....	82
Figure 3.5:	Distribution of number of distinct compounds that are present in different documents .....	83
Figure 3.6:	Empirical Cumulative Distribution Function of observed differences in biological replicate pairs.....	86
Figure 3.7:	Sample MDS plot of several compounds present in a data set.....	102
Figure 3.8:	Schematic View of Proposed Biological Deduplication Workflow .....	108
Figure 4.1:	Typical plot showing improved performance of GA-selected descriptor subset over time .....	125
Figure 4.2:	MDS plot showing relative similarities of $\alpha$ 2a ligand set .....	130
Figure 4.3:	Performance of models of $\alpha$ 2a binding .....	131
Figure 4.4:	Performance of different length subsets of descriptors in regression kNN models.....	132
Figure 4.5:	Compounds showing possible $\alpha$ 2a binding affinity.....	134



Figure 4.6:	MDS plot showing relative similarities of SERT ligand set .....	135
Figure 4.7:	Performance of models of SERT binding .....	137
Figure 4.8:	Commonly occurring fragments from 429 compounds identified as having potential SERT binding affinity.....	138

## LIST OF ABBREVIATIONS AND SYMBOLS

5-HTR <sub>1</sub>	serotonin receptor, type 1
α <sub>2a</sub>	alpha-2a adrenergic receptors
α <sub>2b</sub>	alpha-2b adrenergic receptors
α <sub>2c</sub>	alpha-2c adrenergic receptors
API	active pharmaceutical ingredient
AZ	AstraZeneca
CAS	Chemical Abstracts Service
CCR	correct classification rate
CGI	Common Gateway Interface
CNS	central nervous system
CoMFA	comparative molecular field analysis
CSV	comma-separated values
DARPA	Defense Advanced Research Projects Agency
EMA	European Medicines Agency
EMBL-EBI	European Molecular Biology Laboratory – European Bioinformatics Institute
EU	European Union
FDA	Food and Drug Administration (USA)
GMP	good manufacturing process
GPCR	g-protein-coupled receptor
HTS	high-throughput screening
IMIM	Hospital del Mar Medical Research Institute, Barcelona
IUPHAR	International Union of Pure and Applied Pharmacology
J	Joule
KEGG	Kyoto Encyclopedia of Genes and Genomes

kNN	k-nearest neighbors
MCS	maximum common substrate
MDD	major depressive disorder
MDS	multidimensional scaling
MIABE	Minimum Information About a Bioactive Entity
MIAME	Minimum Information About a Microarray Experiment
NCBI	National Center for Biotechnology Information (USA)
NCI	National Cancer Institute (USA)
NCGC	National Chemical Genomics Center (USA)
NIEHS	National Institute of Environmental Health Science (USA)
NIH	National Institutes of Health (USA)
NLM	National Library of Medicine (USA)
PDSP	Psychoactive Drug Screening Program
RDBMS	relational database management system
REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
rf	random forest
RN	CAS Registry Number
RSC	Royal Society of Chemistry (UK)
QSAR	quantitative structure-activity relationships
SAR	structure-activity relationship
SDF	Structured Data Format
SERT	serotonin reuptake transporter
SMILES	Simplified Molecular-Input Line-Entry System
SPR	surface plasmon resonance
SQL	structured query language
SSRI	selective serotonin reuptake inhibitor
SVM	support vector machine

UCLA      University of California at Los Angeles  
UNC        University of North Carolina  
WOMBAT    World of Medicinal Biomolecular Activity

## Chapter 1: Introduction

*“Everything starts somewhere, though many physicists disagree. But people have always been dimly aware of the problem with the start of things. They wonder how the snowplow driver gets to work, or how the makers of dictionaries look up the spelling of words.”*

Hogfather -Terry Pratchett

Over 2000 years ago, Democratus suggested that the universe was made up of atoms and that the structure of the atoms might be responsible for their properties in obvious mechanical ways. For example, water would be made up of smooth atoms that glide over each other, while iron might be made of atoms with hooks that hold it together and make it strong. While this was a pre-scientific idea arising from an exercise in philosophy, it is nonetheless one of the earliest expressions of a major goal of chemistry: explaining the properties of substances in terms of their atomic and molecular structures and being able to possibly design new substances with desired substances by carefully arranging atoms in molecules.

By the time of Lavoisier and Dalton, much of this grand goal was a distant dream while modern atomic theory was developing and the question of what everything was actually made of came to the forefront. But as organic synthesis was shown to be possible by Wohler (and then profitable by Perkin) the desire to rationally understand (and then predict) the properties of compounds in terms of their structures reemerged. Less than 20 years after the commercial introduction of mauvine and the birth of large scale organic chemical production, Alexander Crum Brown and Thomas Fraser recast Democritus' philosophical musings into more scientific form and proposed that the physiological action

of a substance is a function of the chemical structure<sup>4</sup>. As years passed, experiments demonstrated the merit of this concept, including the correlation of cytotoxicity with decreasing water solubility in organic compounds by Richet in 1893<sup>5</sup>, and the parallel discoveries by Overton and Meyer a few years later that the water/oil partition coefficient also correlated with the ability of organic compounds to induce lethargy and surgical coma<sup>6</sup> (this is the source of the popular rumor that gaseous anesthetics are non-specific drugs that work entirely by disrupting the dynamic flow of lipid-bilayers in the cell membrane<sup>7</sup>).

Simultaneously with this foundational work in chemistry, physicians and biologists were laying the foundations of modern pharmacology by considering how small molecules might interact with other components of biological systems. The work of John Langley in the latter half of the 19<sup>th</sup> century on the contrary physiological effects of atropine and pilocarpine is considered one of the foundational points of receptor biology. By examining how each of those compounds could selectively counteract the effects of the other, Langley conceived of specific receptors in cells which could interact with chemicals with certain properties (in this case, a quarternary amine). This concept was furthered by the advocacy of Paul Ehrlich at the turn of the 20<sup>th</sup> century as he sought specific compounds that would act as antibacterial agents. Ehrlich suggested in his lock and key hypothesis of drug action that it was not merely the presence of certain groups in a compound that would cause a compound to work, but specific structural features of the drug which would mediate interactions with a receptor “like a key turning in a lock”<sup>8</sup>.

Shortly before the outbreak of the Second World War, Louis Hammett was examining the acidity of benzoic acid derivatives and how it related to the rate of amide formation when trimethylamine is introduced into a reaction vessel with one. He noted that there was an almost linear relationship when the acidity (as  $pK_a$ ) when reaction rates were plotted against each other for either the meta- or para-substituted benzoic acids<sup>3</sup>.

Ultimately, this analysis yielded a set of constants relating to the electron donating or withdrawing properties of the substituted group that was invariant across multiple kinds of reactions. This constant, sigma ( $\sigma$ ), captures many of the properties of different substituent groups in a way that mirrors the experience and intuition of working chemists in that substituents with similar sigma values, such as chloro- and bromo- or methyl- and isopropyl- will have similar effects on chemical reactivity when substituted on the same base structure. Equally important, however, is that the sigma constant allows for a congeneric series of compounds to be described numerically in terms of the difference in their structures.

These developments culminated in Corwin Hansch's 1962 publication of a model of the ability of phenoxy acids to promote plant growth by an auxin-like mechanism<sup>9</sup>. This model used the sigma parameter to describe the ability of a given substituent group to activate or deactivate an aromatic ring to further substitution, and a water/octanol partition coefficient to describe a given molecule's ability to penetrate into cells. This led to experiments in predicting small molecule activities in other biological systems and ultimately to the formulation of the Hansch equation<sup>10</sup>, for many years the starting point for much of the work in quantitative structure-activity relationships (QSAR) in biological systems.

The Hansch formulation of QSAR was not without its shortcomings. In particular, it depended heavily on experimental data for not only the activity values, but also for the hydrophobic parameters that it depended upon. The Hansch laboratory at Pomona College was as much an analytical lab that determined the water/*n*-octanol partition coefficient for thousands of compounds of interest as it was a statistical modeling team. Hansch also depended heavily on collaborators in other disciplines to provide assay values for his laboratory to model. Industrial researchers were able to obtain additional consistent assay

values for novel compounds more easily and began to apply the Hansch-Fujita analysis to problems of drug discovery, and new computational tools began to offer the possibility of calculating hydrophobicities instead of measuring them for novel compounds. These datasets remained small, normally containing no more than thirty compounds (often as few as five or ten) and many of the most interesting ones remained undisclosed by the pharmaceutical companies who collected them. In their view, the synthesis of new compounds and their assaying required a significant investment of time and expense, so it would be foolish to just give the results away. Similarly, the synthesis and testing of a SAR series in an academic laboratory required a relatively wide set of necessary skills and did not match well with the research goals of the majority of academic researchers. A few QSAR datasets were nevertheless assembled and became heavily relied upon for methods development and training new practitioners in the art of model building. QSAR was a technology hobbled by a lack of access to data.

Ultimately, this situation changed because of two technological advances in chemistry. The first has its roots in solid-phase supported syntheses of peptides. These methods required the development of highly efficient, selective reactions that tolerated a variety of substituent groups. In particular, the construction of peptidic amide bonds became well optimized. This advance led to the development of more general combinatorial chemistry techniques, where thousands or millions of compounds are synthesized from combining two series of fragments which both bear one of the two requisite groups for a generic coupling. Starting in the early 1980s, synthetic chemists often were able to use combinatorial techniques to elaborate on an initial lead compound against a given target and quickly create an entire SAR series to attempt the optimization of a new drug or probe.

The second advance which enabled large-scale QSAR was the development of the high-throughput screen (HTS). Much as the desire for efficient chemical synthesis of novel



peptides drove the development of more general combinatorial chemistry methods, the transition of the pharmaceutical industry to target-based approaches in the 1970s, coupled with the advances in robotics first seen in heavy industry around the same time, allowed biological assays to move from being labor intensive and tissue- or organism-based to being constrained by the availability of materials. For radioligand displacement or fluorometric assays that might only require 15 to 60 minutes to run, converting to HTS technologies allowed organizations to increase their throughput two to three orders of magnitude by migrating those experiments to 96- or 384-well plate formats. This transition also simplified collaborations between assay developers and synthetic chemists or computational modelers. Whereas screening a set of compounds against a target might have been several day's work for a technician or student, there was little more effort involved in setting up a screen for 50 or 100 compounds than there was in setting up for a single compound. Eventually, funding would become available for shared screening facilities to be developed, providing assay services for chemical biologists in the non-profit sector at low marginal cost to their research programs. The trade-off for this low cost, however, is usually that the assay results must be made available to the general public. Further advancing the general availability of chemical data have been projects such as ChemSpider<sup>1</sup>, and the purchase and conversion to public access of the StARlite database by the Wellcome Trust and EMBL-EBI.

The increased availability of HTS to the broad base of researchers, coupled with the ensuing surge in chemical activity data has allowed chemistry to become part of the era of Big Data and data science. "Big Data" has become something of a buzz-word since its coining sometime around 1990. It is a shorthand term which encompasses the increasing tendency of raw data of multiple forms to be stored in a computer-readable format that can then be winnowed, sorted, and recombined to hopefully identify trends and patterns which no single human could manually identify or extract. The key aspects of big data vary from source to source, but there seems to be a general agreement that for something to be in the realm of

big data, the source material needs to be in electronic form that is at least an order of magnitude more than can be reasonably handled in the core memory of the largest available computers and that the analysis needs to be computationally complex enough that it necessitates the use of parallel computing techniques to complete in a reasonable time scale<sup>11</sup>. A no less incomplete, but arguably more authoritative definition was recently given by Phillip Bourne, the director for big data projects at NIH, entitled “What Big Data Means To Me.” In it, Bourne suggests that Big Data is an emergent property of organizations which have transformed their operations into a truly digital enterprise and moved all their functional data into electronic storage, allowing correlation of that data to make new connections and gain new insight into their core processes<sup>2</sup>. He cites the example of a hypothetical university where all laboratory data is stored in an electronic repository, and a chemistry graduate student’s research into an organism being used for bioremediation of toxic waste is correlated with an MD/PhD student’s research notes on neurodegenerative diseases by virtue of having several genes annotated as being of interest in both data sets, thus leading to a connection, collaboration, and small-molecule analog of that gene’s product showing promise in pre-clinical testing. While this is certainly a best-case scenario, it certainly identifies the enthusiasm felt by many of the proponents of Big Data as an approach to science.

The changes that have allowed chemistry to become a big data-driven undertaking follow, but mirror, the evolution of biology into a data-driven, mechanistic discipline. In biology, these changes can be traced to the determination of pioneers in gene sequencing and molecular genetics to ensure that their research results would not be locked up in printed formats and/or private data silos, but rather would be freely exchangeable in the spirit of peer-reviewed research. To this day, it is almost uniformly required that new genetic or protein sequences be submitted into public repositories before they can be discussed in the peer-reviewed literature. Similarly, researchers are expected to make available vectors

for new sequences and constructs as a condition of publication. In crystallography and expression analysis, other public repositories fill analogous roles. The publication of a manuscript is not a right conferred by the willingness to pay page charges, but rather an accession to cooperative mutual exchanges with the larger community.

Data science and Big Data are popular topics with pundits, futurists, and seers right now, but they do not represent an effortlessly sunny future for researchers, bedecked with rainbows and populated by friendly unicorns. Big Data cannot magically overcome all the mundane problems that are currently being seen in biomedical research, however. In 2011, researchers at Bayer examined 67 completed projects, and found that it was impossible to replicate the findings of almost two-thirds of them<sup>12</sup>. Likewise, a survey of high-impact journals in cancer research by a group at Amgen in 2012 found that attempting to replicate the findings of fifty-three high-profile reports failed in forty-seven cases<sup>13</sup>. The question of data quality and reproducibility is serious enough that Francis Collins, the director of the NIH, published an editorial suggesting that some new projects in translational medicine may be required to show independent validation of results before being used as the basis for further funding proposals<sup>14</sup>. It seems obvious that there is a problem with data reliability when a peer-reviewed journal offers an editorial entitled “Reproducibility: Six Flags for suspect work” which offers exactly that, a list with explanatory text that offers to help researchers know when the results in a report are potentially too good to be true<sup>15</sup>.

One might think that, as a computational discipline, bioinformatics would be more resistant to errors. Unfortunately, this is not the case. In 2009, Phillip Bourne’s group published a paper suggesting targets in tuberculosis that might be susceptible to interacting with already approved drugs for other, unrelated conditions<sup>16</sup>. Two years later, he revisited the workflow for this project with a group of computer scientists who had no particular background in computational biology<sup>17</sup>. Their goal was to recreate the workflow and results

described in that paper with only that paper as a reference. This attempt failed spectacularly, with only one out of fourteen distinct steps in the workflow successfully recreated. The restrictions on the researchers attempting to reproduce the project were then loosened, and they were allowed to consult other resources: first, the documentation for various programs being utilized, and second, the researchers who actually did the research described in the original paper. After checking external documentation, only half of the fourteen workflow steps were successfully reproduced, and after consulting with the experts on the process, all fourteen were functional. However, the results from the workflow did not exactly reproduce the original results. The compounds identified in the original workflow were reisolated, with similar but not exactly identical  $p$  values for significance, and many other compounds that had not appeared in the original paper were identified with acceptably low  $p$  values. In the conclusions, it is proposed that there are multiple sources for these problems, including the relentless addition of new data to public, web-based data repositories, the embedding of significant parameter values in run scripts that are not made part of the manuscript, and using manual editing processes in text editors or spreadsheets over scripted transformations which can be audited and checked as long as the code exists<sup>17</sup>.

Even if the work completed is perfectly documented, this does not ensure its accuracy. A short trip through the computational biology literature reveals multiple kinds of errors that are made, arising from either mechanical error or a basic misunderstanding of the tools available. For example, logic errors in programs from incorrectly mixing 0-offset indexing in arrays and 1-offset indexing arise both because scientific developers make logic errors in coding their algorithms and because they have been crossing back and forth between languages that use different conventions<sup>18</sup>. A recurrent problem stems from the use of Excel as an intermediate data storage, editing, and export tool. In 2004, it was noted in BMC Bioinformatics that datasets with gene names that can be parsed as dates were being inappropriately renamed as those dates; DEC1, OCT4 and SEPT9 were being renamed as 1-

Dec, 4-Oct, and 9-Sept, respectively<sup>19,20</sup>. A solution is to be careful when importing CSV data, to explicitly mark columns with gene names as text, and not to allow Excel to auto-detect date information. This problem is not unusual. It occurs in other forms in other domains, both scientific and non-scientific. In cheminformatics, Excel will often attempt to either parse CAS registry numbers as dates or as subtraction formulas. In spite of this, a decade later, people still find these mangled gene names in supposedly curated datasets. Similarly, a recurrent issue that NCBI deals with is complaints about supposed errors in GenBank records by users who misunderstand the relationship between GenBank and RefSeq (RefSeq is a curated database where incorrect information is corrected, GenBank is an archival database of sequence information as reported in the literature or by direct submission by the sequencer)<sup>21</sup>. An even more basic error comes from novice users who do not fully comprehend that Linux is fully case-sensitive, and so is the grep command by default. Failing to specify a case-insensitive search has repeatedly caused difficulty in workflow scripts that were supposedly production ready, causing relevant data to be omitted from compiled subsets of larger databases.

For practitioners of cheminformatics, these examples are useful cautionary tales. They are, however, not the only evidence that there are problems needing to be addressed in the sources of data we use and the methods we use to handle it once it has been compiled and extracted.

In 2008, Douglas Young and colleagues published a study showing that there were erroneous chemical structures in their own QSAR models, even though they had used reputable commercial vendors as the source of those structures<sup>22</sup>. Furthermore, they determined that there was an error rate of up to 3.4% in those sources, and that having errors at that proportion was enough to cause significant error in their models. Their conclusion was that errors in chemical structure are more common than had been suspected,

and that for the majority of compounds examined, the inaccuracy of the structures resulted in a decrease in the predictability of the compound's toxicity<sup>22</sup>.

Similarly, in 2010, our laboratory showed the need for chemical structures to be standardized and to have any duplicate values removed from a data set being used for QSAR modeling. Errors were identified for about 10% of the structures in the NCI's AIDS antiviral screen, and there were six pairs of duplicated compounds with pIGC<sub>50</sub> values differing by up to one whole log unit in a well-known data set for aquatic toxicity<sup>23</sup>. The ultimate estimate of an acceptable error rate in QSAR data was that anything above 5% would result in models with poor external predictivity and erratic performance in general. In order to identify duplicates and prevent difficulties with descriptor calculation arising from the presence of unsupported atoms, Tropsha, *et al.*, provide a detailed workflow for the handling of chemical structures associated with a QSAR model<sup>23</sup>.

Most recently, Tiikkainen, *et al.*, compared the contents of three major bioactivity databases: ChEMBL, WOMBAT, and Linceptor<sup>24</sup>. ChEMBL is a publicly available bioactivity database based on the formerly commercial StarLite database which seeks to cover as much of the traditional medicinal chemistry literature as possible<sup>25</sup>. WOMBAT is a commercially curated database that seeks to provide highly accurate values for various biological measurements of use in the pharmaceutical industry<sup>26</sup>, and Linceptor is another commercial bioactivity database that seeks to cover as much of the peer-reviewed and patent literature as is possible<sup>27</sup>. Between the databases, they have 5,013,463 activities recorded, but only 1.5% of these (73,076) are actually present in all three sources. Four different values were compared for each shared activity: the ligand structure, the target identity, the value of the reported activity and the type of activity reported. In cases where one source differed from the other two, that source was presumed to be the incorrect one, while in cases where all

three values differed, there was no a priori way to detect which one was correct. These results are summarized in Table 1.1.

	<b>Ligand Errors</b>	<b>Target Errors</b>	<b>Value Errors</b>	<b>Type Errors</b>
<b>ChEMBL</b>	2181 (5.2 %)	1454 (3.2 %)	445 (1.1 %)	9 (0.0 %)
<b>Liceptor</b>	2968 (7.1 %)	1134 (2.6 %)	1072 (2.5 %)	72 (0.1 %)
<b>WOMBAT</b>	2491 (6.0 %)	819 (1.8 %)	510 (1.2 %)	94 (0.2 %)
<b>Discordant</b>	2429 (5.9 %)	214 (0.5 %)	54 (0.2 %)	0 (0 %)

Table 1.1: Raw occurrence of different error types and specific error rate estimates for each data source considered. (After Tiikkainen <sup>24</sup>)

After the initial counts of divergent values, a limited subset of 45 inconsistent activity values were selected from the list of all inconsistencies and manually checked in the primary literature. Of these, 37 (82.2 %) correctly identified the incorrect value as being the discordant entry while in three (6.7%) cases, the majority value was incorrect. In the remaining five cases, the correct value could not be ascertained because of insufficient information in the source. It is worth noting that this method primarily identifies situations where the compilation process for the database or the storage method for the database is introducing errors. This analysis obviously cannot identify when errors were introduced by the initial preparation and publication of the primary manuscript. It is also noteworthy that the represented chemical structures by themselves have an error rate above the 5% threshold identified by Fourches, *et al.*, for the construction of reliable models<sup>21</sup>. Nevertheless predictive models are regularly constructed from the ChEMBL and WOMBAT databases. In many cases, the ability to construct successful models can be attributed in many cases to the fact that many of the structural errors identified in both WOMBAT and ChEMBL are, in fact,

stereochemical errors which would not be noted by conventional 0D, 1D and 2D descriptors (38% of ChEMBL and 60% of WOMBAT).

QSAR models have traditionally been used early in the drug-discovery process, for virtual screening and for assessing the relative merits of alternate analogue series. Researchers further along the pipeline in preclinical development might apply models to predict physical properties, pharmacokinetic and pharmacodynamic parameters, and possible toxicological consequences, but these models are still primarily for the internal use of researchers. This is slowly changing. The REACH registration initiative under development in the European Union seeks to ultimately rely entirely on computational models as the first pass of their hazardous materials classification scheme<sup>28</sup>. Similar programs are under early investigation by the United States Environmental Protection Agency. The US Food and Drug Administration is also exploring ways that chemical models can expedite their own regulatory missions. These are exciting times to be practicing cheminformatics, but these and other initiatives will rapidly fall from grace and burn if the models being produced are inaccurate in their predictions, either because of their foundation on bad data, or for another cause. Developing mechanisms to police the data and prevent bad information from being incorporated into new models is a task that needs to be undertaken now.

There are two primary criteria for data to be useful in almost any technical arena. The data must be in a consistent and useful state, and they must be accurate. In the past decade, machine-readable bioactivity data has become increasingly available to the research community through common repositories, and multiple tools and processes for information standardization have become widespread. Problems with data quality stubbornly persist, with apparent error rates in those repositories exceeding the recommended threshold proposed by the cheminformatics community. While there are currently efforts underway to



reduce these error rates at the compilation stages (*e.g.* some of the additional functionality included in recent ChEMBL releases), there is still a need for an extension of the workflow proposed in Fourches, *et al.*, that addresses issues with biological data, and that provides more explicit guidance for which data points should be removed during the deduplication process. It is the intention of this work to elucidate and index the origin and magnitude of errors found in both chemical structure data (Chapter 2) and bioactivity data (Chapter 3). This information stands as the basis for the aforementioned extension to the existing curation workflow. Finally, these principles are applied to the creation of QSAR models suitable for virtual screening experiments seeking to identify ligands showing affinity for two major targets in the central nervous system (Chapter 4).

## REFERENCES

- (1) Brown, A. C.; Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action; *Journal of Anatomy and Physiology* **1868**, *2*, 224–242.
- (2) Richet, M. C. Sur Le Rapport Entre La Toxicite et Les Proprietes Physiques Des Corps. *Comptes des Seances de la Societe de Biologie et de ses Filiales* **1893**, *9*, 775.
- (3) Meyer, H. Zur Theorie Der Alkoholnarkose Erste Mittheilung. *Naunyn-Schmidberg's Archives of Pharmacology* **1899**, *42*, 109–118.
- (4) Forman, S. A.; Chin, V. A. General Anesthetics and Molecular Mechanisms of Unconsciousness. *International Anesthesiology Clinics* **2008**, *46*, 43–53.
- (5) Silverstein, A. M. *Paul Ehrlich's Receptor Immunology: The Magnificent Obsession*; Academic Press: San Diego, CA, 2002; p. 202.
- (6) Hammett, LP. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. **1937**, *59*, 96-103.
- (7) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1963**, *194*, 178–180.
- (8) Hansch, C.; Dunn, W. J. Linear Relationships between Lipophilic Character and Biological Activity of Drugs. *Journal of Pharmaceutical Sciences* **1972**, *61*, 1–19.
- (9) Williams, A.; Tkachenko, V. The Royal Society of Chemistry and the Delivery of Chemistry Data Repositories for the Community. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 1023–1030.
- (10) Ward, J. S.; Barker, A. Undefined by Data: A Survey of Big Data Definitions, <http://arxiv.org/abs/1309.5821> **2013** (accessed 25 Oct 2014).
- (11) Bourne, P. What Big Data Means to Me. *Journal of the American Medical Informatics Association* **2014**, *21*, 194.
- (12) Prinz, F.; Schlange, T.; Asadullah, K. Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? *Nature Reviews. Drug Discovery* **2011**, *10*, 712.
- (13) Begley, C.G.; Ellis, L.M. Drug Development: Raise Standards for Preclinical Cancer Research. *Nature* **2014**, *483*, 531-533.
- (14) Collins, F. S.; Tabak, L. A. Policy: NIH Plans to Enhance Reproducibility. *Nature* **2014**, *505*, 612–613.
- (15) Begley, C.G. Six Red Flags For Suspect Work. *Nature* **2013**, *497*, 433–434.
- (16) Xie, L.; Li, J.; Xie, L.; Bourne, P. Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network to Explain the Side Effects of CETP Inhibitors. *PLoS Computational Biology* **2009**.

- (17) Garijo, D.; Kinnings, S.; Xie, L.; Xie, L.; Zhang, Y.; Bourne, P.; Gil, Y. Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome. *PLoS ONE* **2013**, *8*, e80278.
- (18) Leipzig, J. Question: What Are The Most Common Stupid Mistakes in Bioinformatics? <https://www.biostars.org/p/7126/> (Accessed 24 Oct 2014)
- (19) Zeeberg, B. R.; Riss, J.; Kane, D. W.; Bussey, K. J.; Uchio, E.; Linehan, W. M.; Barrett, J. C.; Weinstein, J. N. Mistaken Identifiers: Gene Name Errors Can Be Introduced Inadvertently When Using Excel in Bioinformatics. *BMC Bioinformatics* **2004**, *5*.
- (20) Saunders, N. "What You're Doing Is Rather Desperate". Gene name errors and Excel: lessons not learned. <http://nsaunders.wordpress.com/2012/10/22/gene-name-errors-and-excel-lessons-not-learned> . (accessed 24 Oct 2014)
- (21) Moran, L.A. Errors in Sequence Databases. <http://sandwalk.blogspot.com/2008/06/errors-in-sequence-databases.html> . (accessed 25 Oct 2014)
- (22) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science* **2008**, *27*, 1337–1345.
- (23) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50*, 1189–1204.
- (24) Tiikkainen, P.; Franke, L. Analysis of Commercial and Public Bioactivity Databases. *Journal of Chemical Information and Modeling* **2012**, *52*, 319–26.
- (25) Gaulton, A.; Bellis, L.; Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research* **2012**, *40*, D1100–1107.
- (26) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T.I. WOMBAT: World of Molecular Bioactivity. in *Chemoinformatics in Drug Discovery*; Wiley-VCH: New York **2005**.
- (27) Evolvus, Ltd. LICEPTOR Database. <http://liceptor.com/products/databases/liceptordatabase.html> . (accessed 25 Oct 2014).
- (28) Worth, A. P. The Role of QSAR Methodology in the Regulatory Assessment of Chemicals. In; *Challenges and Advances in Computational Chemistry and Physics*; Springer Netherlands: Amsterdam, 2010; pp. 367–382.

## Chapter 2: Chemical Structure Accuracy

*“Do not consider it proof just because it is written in books, for a liar who will deceive with his tongue will not hesitate to do the same with his pen”*

Mosheh ben Maimon

### Summary

It has been established for some time that chemical structures occurring in public and private chemical data sets are not 100% accurate. Multiple estimates have been made of the fraction of incorrect structures in large data sets ranging from 3.4 to 10%. This is particularly a problem for QSAR modeling, as it is also known that having more than one incorrect structure in twenty is sufficient to undermine the accuracy of a QSAR model. As bioactivity databases become larger and more accessible, it becomes possible to construct QSAR models with data for more compounds from more sources. This makes manual verification of chemical structures a slower and more tedious process, and also increases the risk that a functional group may be inconsistently represented in two or more structures, which is as detrimental as an outright structural error to the quality of the overall model. The goal of this study is to compare the methods that modelers might use to assign structures to a list of named chemical compounds and then to compare the accuracy of several different sources that might be used by chemists to resolve names to structures.

Initially, a list of non-systematic names of well-defined chemical substances was needed. Small molecule drugs were chosen as search targets because they are well described in both scientific publications and regulatory filings, possess clear (but arbitrary) names, and are of relatively broad interest. A list of top-selling drugs from 2006 was obtained and filtered to yield 151 generic names. Four groups of cheminformaticians independently

devised workflows to find the chemical structure corresponding to each name, and retrieve that structure in machine and human readable formats. Two groups developed manual consensus search methods that directly searched several well-known structure repositories (one with manual comparison and transfer of structural information, and the other using automated comparisons and transfer in machine-readable formats). One group made automated searches of a pre-existing, pre-curated internal structure database which was developed, in part, using many of the same sources that the previous two groups searched. The final group used an automated live search of Internet resources with consensus consolidation of structures to find the structure associated with a given name. This yielded four sets of 151 chemical structures. Representatives from each team discussed any mismatches between structures for each compound, and a consensus gold list was finalized reflecting the best structure identified for each compound. This gold list was then used to search several public chemical structure databases. For each of these databases, the number of hits returned for searches on each name and the number of times that searching by a name returned the same structure as the gold list were returned. Finally, incorrect structures from both the initial gold list assembly and searches of public databases were examined to determine what kinds of errors were predominant.

After the assembly of the gold list, each group's accuracy as compared to the final gold list was evaluated. No team correctly identified all 151 chemical structures and there was an initial consensus of opinion for approximately 75% of all compounds. The two groups which combined manual curation with automated matching and structure handling had the best performance with 3 and 6 incorrect structures. The UNC group which relied more heavily on visual comparisons and manual structure comparisons had 14 incorrect structures, while the fully automated Internet search group had 30 structures in error. Statistical testing found a significant difference between the fully automated results and all other results, no significant difference for the automated structure handling groups, and a

marginal difference between the manual handling group and one of the two automated structural handling groups. Comparing the gold list to five major sources for chemical structure information, no single source had all 151 structures correct. The accuracies of the sources ranged from 75 to 93% compared to the gold list; three of the five sources returned more than one structure for at least some of the chosen names. Two problem areas were also identified in the software used for some portions of the workflow that negatively affected one group's results.

This work is significant because it places some quantitative limits on the performance of manual and automated chemical structure resolution when multiple sources exist. Because Internet data sources are not annotated and curated to a uniformly high level of accuracy, fully automatic searches for chemical structure information are unlikely to return structures of high enough quality for meaningful QSAR modeling. We reaffirm the need for a manual structure curation strategy in QSAR model development, and show the desirability of strategies that combine computational structure handling with human pattern recognition to yield the highest quality data sets for modeling. Finally, we provide an illustration of the advantages of MIABE and other data standards that allow continuous electronic propagation of chemical structure data, thereby eliminating the errors introduced by manual reentry of chemical structures.

## **Introduction**

Molecular structure has been a defining issue in chemistry as long as the discipline has existed. Many of the greatest successes in organic synthesis were motivated by the desire to prove the proposed structure of a natural product correct by reproducing it *in vitro* and demonstrating that compounds from both sources have identical properties. This need for accuracy extends beyond the laboratory. Any situation where chemical structures are used to define the identity of a compound will perforce be concerned with the accuracy of the

structures presented. Beyond utilitarian concerns about chemical structures, generating and using accurate structures is a matter of professional pride and competency for chemists.

Graphical chemical structures have their antecedents in the early 19<sup>th</sup> century work of Dalton and Thompson and began to be standardized between 1860 and 1871 by groups of European chemists<sup>1</sup>. By the time of Gilbert Lewis and his electron configuration structures in the early 20<sup>th</sup> century, organic chemists had informally agreed on conventions for the presentation of structural information that are the antecedent of IUPAC's 2008 Graphical Representation Standards for Chemical Structure Diagrams<sup>2</sup>. While Gmelin's *Handbook* predates this era (beginning publication in 1817) and Beilstein's *Handbuch* began publication immediately after the first standards emerged in 1881, they undertook to ensure that their data sources provided structural information along with names, physical properties and reactions of listed compounds<sup>3</sup>. Similarly, when the American Chemical Society began publication of Chemical Abstracts in 1907, structural information was included for compounds appearing in the chemical literature. The effort to maintain structural information about known compounds took a leap forward in 1965 with the introduction of CAS Registry services that provided unique, externally-visible identifiers for chemical substances<sup>4</sup>.

All of this structural information, however, was accumulating almost entirely in printed form and with few efficient options for structural searches. Attempts were made to enable structure-based searching of the chemical literature, including the CAS Ring Systems Index that classified compounds based on the structure of their topological backbones and the use of systematic nomenclature to index compounds by name. These lacked, however, an efficient way to go from structure to common name or to find other compounds similar to a given compound. While a faithful copy of the structure might be obtained by photocopying, including a structure in a new document almost inevitably required duplicating the

structural diagram manually, with the inherent risks of errors accumulating during that process.

Since the first representation standards were propagated from the first International Chemical Congress in Karlsruhe in 1860<sup>5</sup>, significant efforts had been made in representing chemical structures in typographic form. While these attempts presented an easy mechanism of including chemical structures in typeset documents without the expense of engraving a separate plate for chemical structure diagrams, they were suboptimal because of the limited ways that stereo and regio-isomers could be represented. Further, it was difficult to visualize ways in which topologically distant parts of a complex molecule might become geometrically close and interact. By 1950, there were several groups working on better methods to store chemical structures as alphanumeric strings including, most notably, William Wiswesser, who developed the eponymous Wiswesser Line Notation<sup>1</sup>. This, and several other, encodings allowed chemical structures to be stored easily as alphanumeric strings on punched cards and be searched and collated by IBM's ubiquitous Hollerith technology and, more significantly, using the ever-growing installed base of electronic computers.

The widespread deployment of mainframe computers, combined with the demands of maintaining card indices for the expanding number of chemical structures appearing in the literature, led CAS to migrate to computerized databases as part of the roll-out of their REGISTRY service in 1965. Further developments in search technology, and the growth of commercial dial-up telecomputing providers, led CAS to allow information specialists remote access to their registry through a customized graphical terminal. They subsequently collaborated with FIZ Karlsruhe by launching STN International in 1983<sup>4</sup>. By 1990, STN Express had replaced customized terminals and complicated text queries, allowing non-specialists to search and access chemical structure information from any modem-equipped



personal computer. While access was slow by current standards and very expensive (individual searches could run into hundreds of dollars of access fees), chemists could draw a chemical structure and determine if it had been reported before, without having to meticulously convert the structure to a systematic name or learn complicated line-notations for the search.

Cotemporally, the Internet, a packet-switched data network originally developed by BBN, UCLA and Stanford University under contract for DARPA in the late 1960s, appeared. While it was originally envisioned solely as a test-bed for robust messaging technologies, researchers at major universities, government agencies, and a few fortunate corporate laboratories quickly discovered the wonder of near instantaneous communications and remote access between distant computer facilities. Beyond computer networking researchers, some of the earliest adopters of this technology were molecular biologists and others interested in what has evolved into bioinformatics. They used the new technology to share genetic and protein sequence information, bypassing the need to manually rekey information from publications or to generate portable storage media and ship them to different locations.

When the US National Science Foundation removed their restrictions on commercial use of the NSFnet backbone in the early 1990s, Internet access and utilization exploded. By 1995, the National Library of Medicine (NLM) had placed their *Index Medicus* and PubMed abstracts services online, and CAS followed suit. A major difference was that the NLM offered their service for free (eliminating the expensive requirement of multiple phone lines and banks of modems) while CAS still charged by the search. By that time, the Protein Data Bank held 3814 structures and GenBank held 425,000,000 bases in over 600,000 distinct, identified genes. Some websites showed chemicals structures by embedded graphics and private collections of chemical structures existed in machine-readable form. However, there

were few tools for searching and no publically available, free, central repository for chemical structures.

This changed by the beginning of the 21<sup>st</sup> century, due to the efforts of Peter Murray-Rust, Henry Rzepa and many others who had observed the effect that open data access had on the development of bioinformatics. They advocated for better standards for chemical information storage on the Internet and easier access to that data <sup>6</sup> . With the founding of Wikipedia in 2001, the wiki concept of user-editable content on a web site was applied on a new scale and the idea of crowd-sourcing information collection came to the public consciousness. Of particular interest this discussion, in their quest to collect as much of human knowledge as possible, Wikipedia authors and editors had begun adding machine readable forms of chemical structures to the pages of some compounds. With the ubiquity of web search engines such as Google, AltaVista, and Jeeves, the idea of searching the Internet directly for chemical structure data became feasible.

By 2004, the number of machine readable chemical structures available on the Internet had begun to grow exponentially. The first major source to appear was PubChem, funded as a joint project by the NLM and the NIH Molecular Roadmap Libraries Program in 2004 <sup>7</sup> . While initially intended as a chemical structure repository to support dissemination of the results of HTS assay screens against novel targets by academic groups, it incorporated data from several other NIH sources, including structures referenced in published papers covered by the NIH's open access requirements.

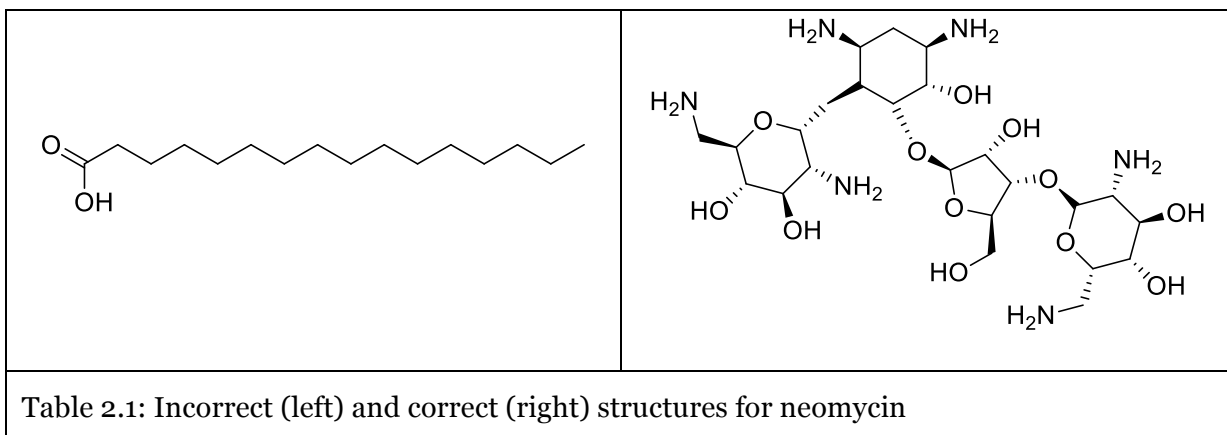
Drawing on the release of PubChem, the increasing amount of chemical structure information on Wikipedia, multiple electronic catalogs from chemical suppliers, and several smaller collections, ChemSpider was founded in 2007 as an experiment in applying crowdsourcing methodologies to scientific data <sup>8</sup> . Any interested person was invited to create an account, submit novel chemical structures with names and any other available

data, and correct or validate the work of other users. The database grew rapidly and began a partnership with Wikipedia where chemical structure data would be automatically retrieved from ChemSpider for display on the appropriate Wikipedia pages. The Royal Society of Chemistry acquired the service in 2009, and while still supporting crowd-sourced curation of chemical structures, ChemSpider expanded into other facets of open science. These aspects include archiving chemical spectra and synthesis methods, providing automated structure validation services and providing chemical data for the OpenPHACTS project, an EU project aiming to provide better quality structured pharmacological information to researchers as an accelerator for drug discovery<sup>9</sup>.

In 2008, the Wellcome Trust acquired the rights to a structure-activity database, StARlite, from Galapagos NV and donated the data, in turn, to the European Bioinformatics Institute<sup>10</sup>. This database abstracted bioactivity information from the published literature (in particular the medicinal chemistry literature as represented primarily by the *Journal of Medicinal Chemistry*, *Bioorganic and Medicinal Chemistry* and *BOMC Letters*). Each compound appearing in conjunction with a bioactivity value was encoded in a machine-readable format and stored in a relational database with the reported value, a citation and other relevant information. Since the original release in 2009, new values and structures from an increasing number of journals have been added, as well as HTS screening results from DrugMatrix, multiple industrial screening programs against malaria and tuberculosis, and other specialized target campaigns.

Not all structures are equally good, however. In April of 2011, the National Chemical Genomics Center of the NIH released a specialized chemical information browser that connected to a new chemical database called the NCGC Pharmaceutical Collection. This database, which combined biological and chemical structure information from NIH, FDA, and private sources, was intended to provide information about all approved active drug

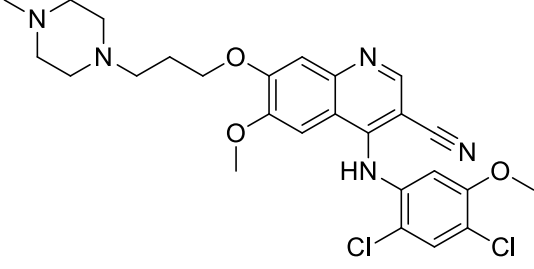
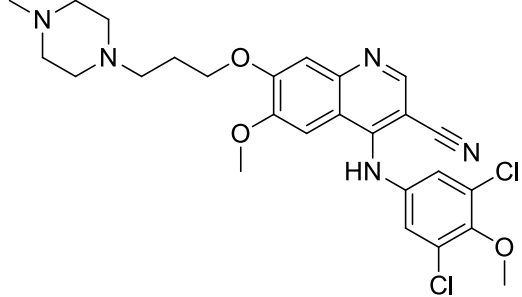
ingredients in the United States as a service supporting drug-repurposing projects. About 2,750 small molecule drugs were included in the release. The launch was conducted to significant public fanfare including a press release with a quote from NIH director Francis Collins, waves of announcements on social media and chemical blogs, and publication of a paper in *Science-Translational Medicine*<sup>11</sup>. Unfortunately, in a blog entry dated April 28 of the same year, Tony Williams noted multiple errors in the data (including bad charge balances, missing and incorrect stereochemistry, and valence errors) from a preliminary scan of the data set<sup>12</sup>. Those errors could be detected without close comparison of the structures to known correct standards. An extended discussion ensued over these and other issues (such as the annotation of stearic acid as one of the six structures given for neomycin as shown in Table 2.1). While an improved database was released relatively quickly with cooperation from Williams and other members of the community, the collection's reputation was already tarnished. The admission by the NCGC team that structure curation was an afterthought assigned to a junior team member did not help their credibility<sup>13</sup>.



While such an event described may be embarrassing to those involved and their superiors, uncertainties about chemical structures can lead to bigger problems. In 2012, in the course of solving the complex structure of Abl tyrosine kinase bound to bosuntinib (a novel kinase inhibitor developed by Pfizer), Nicholas Levinson of Stanford University noted

an anomaly in the electron density of the bosuntinib<sup>14</sup>. Examination of another group's structure of bosuntinib bound to a different kinase showed that, similarly to his own work, the electron density seemed to be at 3 position of a substituted phenyl ring (which is supposedly unsubstituted) rather than at the 2 position (where a chlorine atom is supposed to be, contributing significant electron density). Upon closer examination, his worst fears were confirmed. The material that the chemical supplier sold as bosuntinib was, in fact a regioisomer where the 2-chlorine and 3 hydrogens on the pendant phenyl ring were exchanged and, in addition, another chlorine (at the 4-position) and a methoxy group (at the 5 position) were also exchanged (Table 2.2). The two laboratories had purchased their supplies of the ligand from two different suppliers. Further investigation determined that, of 16 companies claiming to sell bosuntinib, only 2 of them in fact were providing that compound and not the regioisomer. Given that the compound used in the earlier structure was purchased in 2006, there is uncertainty over which compound has been used in experiments since at least that time. With 100 PubMed hits and 300 SciFinder hits, there are many papers which stand to have incorrect structures in them. It appears that the regioisomer is still a (less effective) kinase inhibitor, which limits the impact of this particular error, but the potential impact of mis-mapping a name and structure in another case like this one is large indeed.

Problems with chemical structure accuracy can also have financial consequences. In 1998, Lilly was issued Canadian Patent 2,163,446, a combined design and utility patent covering the use of sildenafil and analogs to treat erectile dysfunction. In this patent, Lilly makes a total of 27 claims regarding these compounds, but only one chemical structure was provided, a Markush structure with 4 substitution positions. All of the claims referenced a dizzying set of options for each substitution (one justice estimated that there were 260 quintillion compounds in these first claims). In the discussion of the claims, Lilly provided a

	
bosuntinib	NOT bosuntinib
Table 2.2: Correct (left) and incorrect (right) structures of bosuntinib	

list of multiple combinations of R groups that were most preferred for pharmaceutical use in erectile dysfunction. But nowhere in the patent did they explicitly identify the structure of sildenafil, the approved and marketed compound<sup>15</sup>. In 2012, the Supreme Court of Canada ruled that this patent was invalid because of this omission, which violates the requirement that a patent provide sufficient information about an invention for it to be reproducible by other craftpersons<sup>16</sup>. While the court reversed themselves on technical grounds regarding standing a year later, the core principle of full disclosure was upheld. This ruling would have opened the door to the marketing of generic sildenafil in Canada two years sooner than would have been the case if the patent had been upheld. While this single case cannot be held up as the source of Lilly's financial ills, the cost of litigation and uncertainty over the forecasting of future sales certainly contributed to their bleaker revenue outlook and depressed share price.

In all three of these cases, there was uncertainty about the relationship between a named chemical compound and its structural formula. The impact on the parties closest to the ambiguity varied, but in all three cases, the structural information and ambiguous designation was available to the general public in machine-readable formats. In the first case, the express goal was to provide structures to the research community. In the second

case, the erroneous structure was given in the published version of multiple scientific papers, but also in electronic catalogs provided by vendors of the misidentified regioisomer. In the last case the patent information, including reprints of the actual application, is freely available from the Canadian Intellectual Property Office. There are almost no safeguards to prevent a public-minded Internet user from taking the (incorrect) information present in any of those sources and adding it to Wikipedia or ChemSpider or even just putting it in a post on their own blog with an InChI key or other text-only representation.

As previously discussed, QSAR modeling is very sensitive to data inaccuracies. A combined biological and chemical error rate of 5% is enough to reduce the performance of a model sufficiently that it will not pass acceptance criteria or, in extreme cases, can prevent the model from converging to a solution at all<sup>17</sup>. With the increase of readily available data, web services such as Chembench and OCHEM, and regulatory agencies relying more and more on computational models for initial screening, there are more non-specialists who will download datasets for a given target or endpoint and attempt to build a model. With no need to manually recode the structures being used, there is a strong temptation to not even look at the chemical structures being used and assume that if the computer can read it, it must be correct. Obviously this represents a risk to at least themselves in wasted time and incorrect assumptions.

This phenomenon raises the question of how accurate freely available chemical structures on the Internet are. Multiple confounding factors make this seemingly innocent question complicated. First, there is no master index for the Internet. Even Google only gets the parts that are not behind unlinked CGI queries and that have not been posted as inhospitable to webcrawlers (via a spiders.txt file). Furthermore, not all structures are created equal. Some of them, like urea or benzene are well known and have been studied over a long period of time; others are from recently synthesized polymers or novel natural

products. It would be difficult to manually assess the accuracy of a structure that has only been reported once without digging deeply into experimental details and spectra.

Assembling a list of chemicals for structural resolution *ab initio* would appear to be a complex problem.

One class of compounds that has well-defined common names and has well-defined chemical structures for almost all members is approved small-molecule drugs. The various government marketing approval mechanisms and good manufacturing processes (GMP) require that the active pharmaceutical components are both well-characterized and that they are identified by names that are systematic (with regards to mechanism of action), yet entirely synthetic. Because this structure/name mapping is of interest to a broader swath of the general population than the chemical literature in general, there are more resources available to the public at large which provide this information, both in print and, either freely or for a fee, electronically. In an ideal world, this mapping would be correct everywhere, for all approved drugs. We recognize this as a naïve assumption and will assume that errors will occur.

While we expect to find errors in the structure/name mapping, we do not know where they will be found. Assuming that one source is always authoritative could lead us to assign an erroneous chemical structure as correct. There is no *ab initio* way to determine how many carbon atoms link the two ring systems in fluphenazine, for example. Rather than attempt to designate one source as completely trusted, having multiple sources of information allows a consensus correct structure can be identified. Having multiple groups of cheminformaticians resolve the names to structure mappings independently, using their own preferred methods, and then comparing the results with each other and allowing discrepancies to be discussed until agreement is reached is more likely to result in the correct structure/name mapping to be found. In addition, this would allow us to compare



how well individual methods for structural resolution perform. Once the searches and comparison of results are complete, the end result is a set of structure/name mappings (referred to as a master list or gold list) whose accuracy we are highly confident in. That list can, in turn, be compared to individual data sources to evaluate the structural accuracy of those individual sources.

In order to test our hypothesis that accurate chemical structural information can be found on the Internet when proper data-handling and curation practices are followed, we joined forces with three other teams of cheminformatics researchers: a group from the Royal Society of Chemistry (RSC) headed by Tony Williams, a group from AstraZeneca headed by Sorel Murtesan, and a group based at IMIM in Barcelona under Jordi Mestres. Each team provided their own preferred workflows with differing amounts of manual intervention and curation. By having each team return a set of chemical structures corresponding to a list of drug names, we can estimate the effect that different search practices have on the accuracy of the structures returned\*.

The objection can be raised that this method is not a perfect replication of the typical QSAR data assembly process where most of the data is taken from analog series in a small number of papers in the public literature or a private data repository. It may, in fact, be more realistic to start with a series of IUPAC names, such as are provided in the experimental details of most medicinal chemistry publications. This would, however, decrease the effect of (or totally eliminate) two sources of errors: the mapping of non-systematic names (such as an internal registry code name, or a proprietary or generic product name) to a chemical structure, and the cumulative effect of structures being copied from source to source (either manually or by automated methods). While it is true (and

---

\* Each group wrote an initial description of their structure resolution workflow which was used as the basis of the descriptions given in the Methods section. Each was modified for consistency in presentation and clarity. All errors and inaccuracies are the fault of the editor, not the original teams.

discussed elsewhere in this work) that most compounds in the public drug discovery literature are only described once, many of the tool compounds which have become ubiquitous in cell biology are known almost exclusively by a trivial name when used and a chemical structure is not commonly provided when the compound is cited; finding the structure for such a compound can be surprisingly complicated. In the latter case, when a structure is copied from source to source, there is a finite (although unestimated) probability that an error will be introduced. This probability, while variable depending on the specifics of how the copying occurs, is incremental; each copy will increase the overall chance that the structure will be perturbed. Using structures that are of more general interest for this survey means that there will be more opportunities for errors to be introduced. In many ways, this work is designed to test for worst-case scenarios in name to chemical structure resolution. The accuracy achieved when data are obtained from a small selection of papers that disclose all chemical structures necessary is likely to be better. These results are useful in their ability to identify effects that would probably be lost in a more homogeneous sample.

## **Methods**

### *Creation of initial drug name list*

The list of drug names to be resolved was obtained from the Wikipedia page “List of largest selling pharmaceutical products” as of 2006<sup>18</sup>. This data was cited as being obtained from the July 2007 issue of *MedAdNews*, although the supposed generic names for each drug had been added by an unidentified editor during the transfer to Wikipedia. These names were actually more of a hindrance than useful because whoever added the information did so inaccurately and stripped out parts of names that were presumed to refer to counterions added in formulation. Crestor became rosuvastatin, not rosuvastatin calcium and Atrovent was listed as ipratropium instead of ipratropium bromide. Unfortunately, this meant that CellSept was called mycophenolate instead of mycophenolate mofetil (removing

information about the ester form actually present), and Co-amoxyclov was identified as amoxicillin, completely ignoring the clavulanic acid component of the drug acting to inhibit beta-lactamases. The list was simplified by removing vaccines, monoclonal antibodies, organometallic compounds, synthetic polymers, and any polypeptide that was composed entirely of the alpha amino acids (in their naturally occurring stereoisomers) that are used in mammalian protein biosynthesis. Proteins containing unnatural residues, including D-amino acids were retained. This yielded a list of 152 drug names which participants were to associate with chemical structures. Premarin and Premapro were initially on the list (as one entry), but were removed due to difficulties in identifying a complete list of active components in those products, leaving 151 entries on the list in total.

For each distinct active pharmaceutical ingredient (API) in the listed drug, teams were to provide a machine-readable structure for the API in the most neutral charge state possible, (in MOL format)<sup>19</sup> with any pharmacologically inactive small ions removed (and with any pro-drug protecting groups removed). An electronic image of that structure represented in standard chemical notation was also requested, along with notes on any oddities or difficulties found in the process of associating a structure with the name.

#### *University of North Carolina Workflow*

Initially, we used ChemIDPlus<sup>20</sup>, DrugBank<sup>21</sup>, ChemSpider<sup>8</sup>, and Wikipedia<sup>22</sup> for establishing the systematic name from the generic names of each compound of the initial name list. These systematic names were used to manually redraw the structure for each compound because of our concerns about the quality of stereochemical information in the different sources. Our initial assumption was that it would be easier to redraw the structures than to identify specific errors, especially in chirality. Usually, the systematic names of the query compounds in the databases were in better agreement than drawn structures. In the

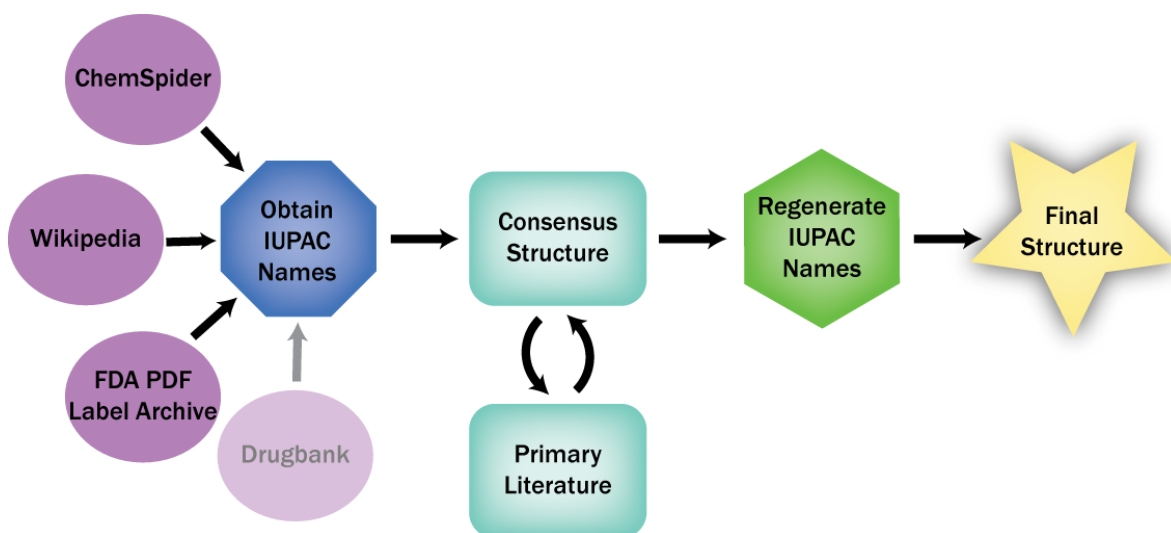


Figure 2.1: Chemical structure resolution workflow as implemented at UNC

most difficult cases (unclear, wrong, or contradictory names) we consulted Google and often found papers in the peer-reviewed literature originating in pharmaceutical companies and initially describing the drug in question, *e.g.* zolpidem<sup>23</sup>. The structures of query compounds were then processed by MM+ force field with gradient < 0.5 as implemented within HyperChem software to provide optimized 3D structure and were manually checked for agreement with the systematic names that we considered correct. Then the resulting ml2 files from HiT QSAR<sup>24</sup> were translated to mol and jpeg formats.

After the first several compounds were completed, it became clear that the workflow described above was not adequate; too much manual effort was required in order to attain the level of accuracy we sought. Furthermore, it appeared that DrugBank and ChemIDPlus conflicted more frequently with other sources and each other than Wikipedia, Chemspider, and the results of Google search.

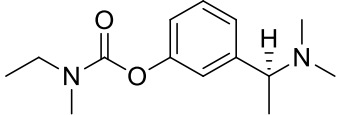
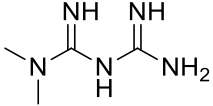
The workflow was then modified to take into account these considerations (as shown in Figure 2.1). For each compound, initial structures were obtained from the structures and systematic names at Wikipedia, ChemSpider, and the FDA approved label archive<sup>25</sup>. The

provided systematic name (usually IUPAC or a recognizable dialect thereof) was converted to a structure, if possible, by the name-to-structure conversion feature of ChemDraw version 11(CambridgeSoft, 2009). These initial and (re)generated structures from each source were visually compared. For each source, the provided chemical structure and the chemical structure derived from the provided chemical name were visually compared. If they were identical, that structure was retained to be compared with the structure obtained from the other sources. Subsequently, the retained structures from the three sources were compared to each other. If there was a discrepancy in the chemical structures provided by different sources, a literature search was undertaken to find the earliest disclosure of the compound. In decreasing order of preference we sought: the earliest disclosure of the compound in the refereed literature, a patent application or patent containing an unambiguous structure (as opposed to a Markush structure), or a crystallographic study assigning an absolute configuration to each stereochemical center. This "literature" structure was then compared to the other candidate structures to determine the most-likely putative structure.

Once the putative structure was determined, the systematic name was generated by the structure-to-name feature of ChemDraw(version 11). This name was immediately back-converted into a structure, and the initial and round-tripped structures were compared. If the structures were not identical, the latter structure was corrected to match the former, and another conversion cycle from structure to name to structure was run. A stable structure to name mapping was reached on a first or second cycle in all but two cases. For these last two drugs, the putative structure was copied into MarvinSketch (ChemAxon, Budapest), and the structure-to-name and name-to-structure functions provided there, were used to generate a stable structure to name mapping. It is worth noting that in these two cases where ChemDraw was unable to converge to a stable structure to name mapping (metformin and rivastigmine) Table 2.1, MarvinSketch was able to interpret the systematic name generated by ChemDraw correctly, but the systematic name that MarvinSketch generated was not

interpretable by ChemDraw. We theorize that this has to do with the relative scarcity of guanadyl and carbamate functionality in daily chemical practice and limited support and testing for their nomenclature. The resulting structures were saved as mol files and exported to PNG format

This workflow can be characterized as manual search, manual curation and manual structure handling.

	
rivastigmine	metformin
Table 2.3: Compounds for which ChemDraw was unable to generate correct systematic names	

#### Royal Society of Chemistry Workflow

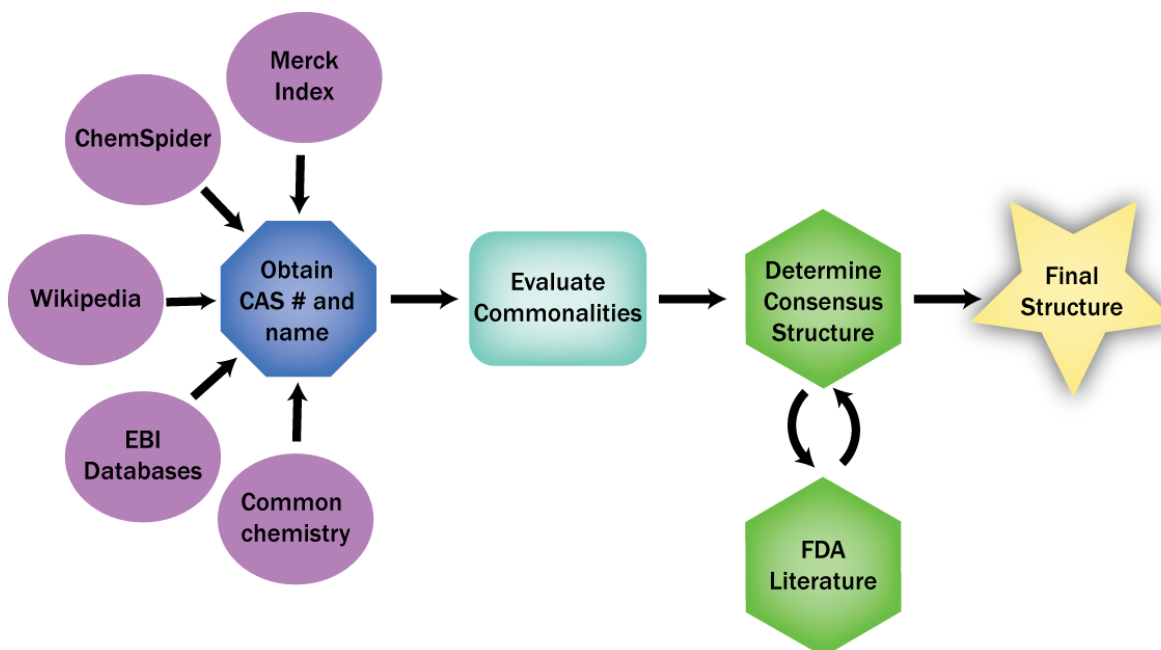


Figure 2.2: Chemical structure resolution workflow as implemented by RSC

The assembly of the dataset was initiated by consulting Wikipedia for an entry with a chemical structure. If a structure was present, it was then assessed to determine whether it was reasonable in terms of whether the article title, the structure image and the body text are referring to the same compound. Possible stereochemistry issues were assessed by checking for how many undefined stereocentres there are. The chemical name from the original file was used as the basis of a search across ChemSpider, ChEBI<sup>26</sup> and the CAS Common Chemistry pages<sup>27</sup>. The CAS Common Chemistry result was the preferred source and consulted initially. If the compound was not found, a CAS Registry Number (RN) was sought from the appropriate Wikipedia article. If present, this RN was used to search Common Chemistry instead of the initial name. If a structure was found, it was assumed to be a good starting point as a basis of searching for contradictions from several other searches.

The results of a name-based search in ChemSpider were then reviewed. Since this database has been actively curated for several years, especially around common compounds and drugs, one compound was returned in general. Occasionally multiple results were returned. These were curated based on an iterative review of multiple sources until only one candidate structure remained.

The initial candidate structure was also checked against the EBI databases ChEBI and ChEMBL. ChEMBL<sup>28</sup> was found to be problematic in terms of the misassociation of synonyms where names would be associated with multiple structures as well as confusion between names for neutral forms versus salts. The Merck index<sup>29</sup> was also used as a reference collection. In those cases where it was difficult to find 2 or more agreeing representations alternative 'authoritative sources' were consulted. These nominally authoritative sources were frequently documents relating to regulatory approvals from the FDA or European Medicines Agency (EMA).

Peptide and peptidomimetic structures were often described as a sequence of amino acids. In these cases the peptide was built using amino acid residues and then used as the basis of structure searching and confirmation against the chemical name. Alternately, the structure presented on Wikipedia was drawn and *in silico* digestion was performed by hydrolysing the peptide bonds. Individual residues were then identified and compared to structures presented on ChemSpider for accuracy.

This resolution scheme can be described as manual search, manual curation and electronic structure handling.

#### *AstraZeneca Workflow*

AstraZeneca Mölndal has developed an internal web service for chemical structure searching and structure/name resolution. Dubbed “Chemistry Connect”, it has integrated 17 internal and external data sources and consolidated over 70 million compound records into 45 million distinct structures. It has been more extensively discussed elsewhere in a paper by the developers, but the key features will be covered here <sup>30</sup>.

Structures being added to Chemistry Connect have their structures standardized according to internal representation standards during the initial registration process. A majority of the compounds are registered as a parent organic structure, although the capacity to include formulation details and mixtures exists for specialized needs. During the registration process, identifiers are sanitized and regularized and included in a large (over 100 million term) dictionary linking structures to names and supporting text-based queries. While there is currently no mechanism in place to force all terms in the dictionary to return the same single chemical structure, there are two mechanisms in place to assist with identifying the correct structure when conflicts occur. First, the order in which structures are returned from a query is tied to the number of independent sources attaching that structure



to annotations matching the search string. This keeps sources with misannotations from being the primary resource returned for a structure to name mapping unless there is no consensus over what the correct structure actually is. Also, a crowdsourcing function has been added to the web service. This allows scientists to modify erroneous records on the spot, instead of filing error reports that need to be processed by a centralized group. While this does not prevent bad results from ever being returned, it reduces the opportunity for obviously incorrect information to remain in place beyond the first time it is noticed.

For this process, each individual drug name was submitted as a query to Chemistry Connect and the first structure returned (in the form provided by Chemistry Connect) was reported as a candidate structure for the gold list.

This resolution scheme can be described as automated search, manual curation and electronic structure handling.

### *IMIM Workflow*

Jordi Mestres' group at IMIM (Hospital del Mar Medical Research Institute, Barcelona), in conjunction with ChemoTargets, SL, a small, early-stage technology company in Barcelona, has developed a non-supervised, weighted strategy for automatic retrieval of chemical structures from free Internet sources given a name or other identifier. Under this method, a master dictionary of drug names and other identifiers was created from KEGG<sup>31</sup>, DrugBank, CheEMBL, IUPHAR-db<sup>32</sup>, PubChem<sup>7</sup> and Wikipedia. Using this thesaurus, a comprehensive search set of identifiers for each drug occurring on the initial search list was created. Each comprehensive set of identifiers was then used to search an undisclosed set of web sites for chemical structures. This yielded a set of chemical structures associated with the original drug name searched, clustered by the identifier used to make the association. A structural search of each returned set was used to identify other synonyms for the original

drug name not occurring in the original dictionary. In the event that there was overlap above an undisclosed threshold between the structures of two different synonym sets that were also linked by a number of latent synonyms, the two sets were merged into a single set.

The final assignment of a single structure to the synonym set associated with a single drug name was accomplished with an unsupervised, weighted vote scheme. Within each synonym set, each structure present was assigned a consistency score, which was proportional to the number of sources including the association between that structure and the drug name being searched for, and to the relative confidence the developers had in the accuracy of the structures from a particular source. The structure with the overall highest consistency score was then sent as a candidate structure for the gold list with no additional visual inspection or manual intervention.

This resolution scheme can be categorized as automated search, automated curation, and electronic structure handling.

### *Structural Comparisons*

Resolved structures were converted with the reference implementation of the InChI Trust software<sup>33</sup> into InChI strings and InChI keys and placed into a central repository. These strings were matched for literal equality to determine whether or not a consensus structure had not been reached. For compounds where there was a discrepancy, participants were asked to provide documentation of their rationale for the structure that was returned and any thoughts about the perceived quality of that structure. Discussion continued around these points until a consensus structure was agreed upon. The gold list of agreed-upon consensus structures was generated in ACD/ChemSketch (version 11 ACD/LABS, Toronto), and saved and distributed to participants as a SDF-formatted file.

### *Statistical Methods*

Fisher's exact test as implemented in Prism (version 6, GraphPad Software, San Diego CA) was used to determine the significance of the differences in error counts between each group's search algorithm.

## Results

The list of initial names as taken from Wikipedia and adjusted before distribution to the participating teams, can be found in Appendix 1.

The final gold list of structures generated by consensus can be found in Appendix 2.

The number of correct structures in each team's results, relative to the final gold list, are found in Table 2.4.

	<b>Total Correct</b>	<b>Stereochemical Errors</b>	<b>Percent Correct</b>	<b>IMIM</b>	<b>AZ</b>	<b>UNC</b>
<b>RSC</b>	148	1	98.0	*		*
<b>UNC</b>	137	5	90.7	*		
<b>AZ</b>	145	2	96.0	*		
<b>IMIM</b>	121	14	80.1			

Table 2.4: Summary of search results by each team compared to the consensus gold list (151 total compounds). An asterisk in the last three columns indicates a statistically significant difference between the two groups in the number of correct structures found ( $p \leq 0.05$ , Fisher's exact test).

### *Evaluation of chemical structure repositories*

A summary of the results from comparing several Internet sources for chemical structures to the structures in the gold list can be found in Table 2.5 and Figure 2.3. Each source of chemical structures was searched by the names used to construct the gold list. The first row and left side of each bar give the total number of structured returned for all names;

a value greater than 151 indicates that one or more name to structure searches returned more than one structure. The second row and right side of each bar represent the number of times the returned structure or structures from each search contained the same structure as was in the gold list. Ideally both numbers would be 151, the number of compounds in the gold list. If the number of correct hits is low, then the chemical structures associated with a data source are not particularly accurate. If the ratio of total hits to correct hits is large, there are multiple structures associated with each name in that data source. The correct structure may be present, but a simple search by name will require that the end user will need to have a way to disambiguate the correct structure, if present, from the other structures returned. A well-curated source will make clear which structures are returned by a name search, and provide a mechanism to disambiguate the structures returned if that is anticipated as a default use case.

<b>Source</b>	<b>NCGC</b>	<b>NCI</b>	<b>PubChem</b>	<b>DrugBank 3.0</b>	<b>DrugBank 2.5</b>
<b>Total Hits</b>	243	142	224	148	150
<b>Correct Hits</b>	113	115	141	117	118
<b>Ratio</b>	2.15	1.23	1.59	1.26	1.27

Table 2.5: Number of hits returned for compounds on consensus gold list for different open access structural databases, and number of correct hits returned.

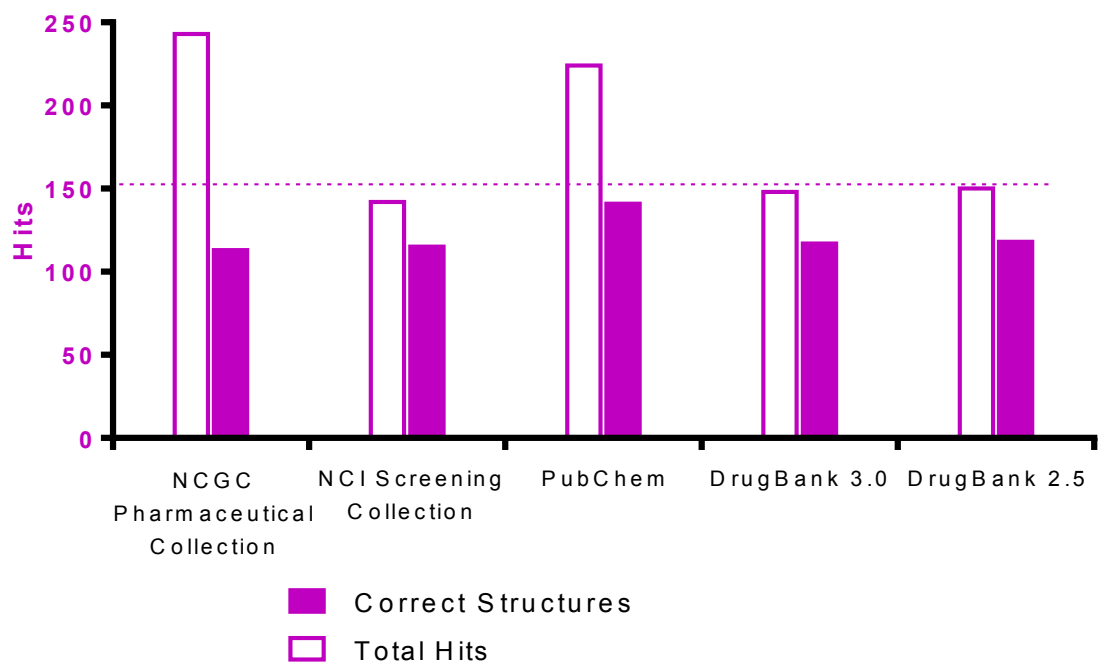


Figure 2.3: Graphical comparison of performance of different sources against consensus gold list. The dotted line is at 151, the total number of compound names searched. Bars that lie closer to that line represent better outcomes than those terminating further away.

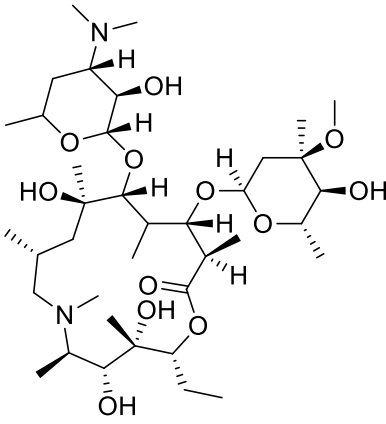
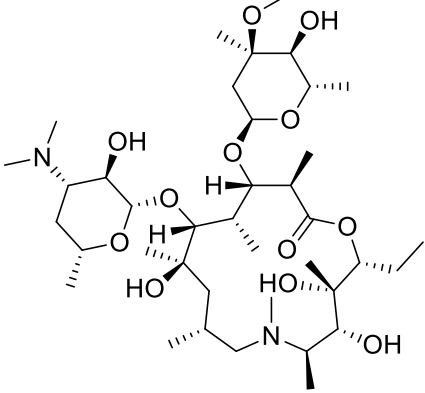
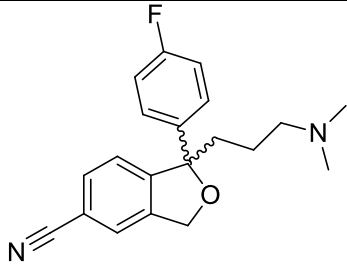
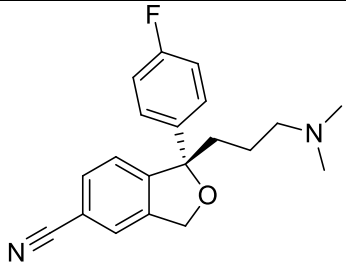
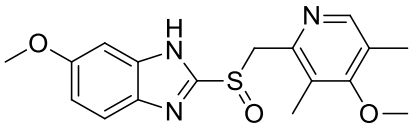
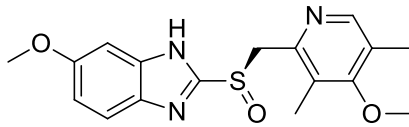
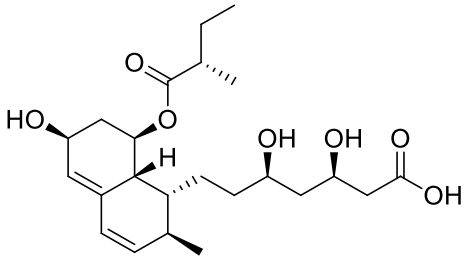
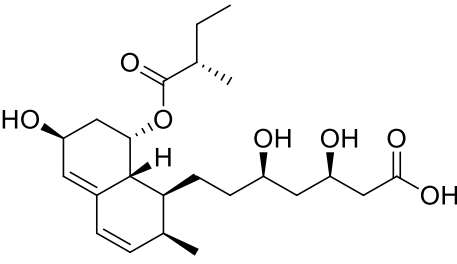
Name	Structure Presented	Correct Structure
Azithromycin	 <p>The structure shown is a complex macrolide with multiple hydroxyl groups and methyl groups. It features a 15-membered macrolide ring with a nitrogen atom at the 14-position. The structure is drawn with a mix of wedge and dash bonds, but the stereochemistry is not clearly defined in a way that matches the correct structure.</p>	 <p>The correct structure of Azithromycin is a 15-membered macrolide ring with a nitrogen atom at the 14-position. It has a methyl group at C-1, a methyl group at C-2, a methyl group at C-3, a methyl group at C-4, a methyl group at C-5, a methyl group at C-6, a methyl group at C-7, a methyl group at C-8, a methyl group at C-9, a methyl group at C-10, a methyl group at C-11, a methyl group at C-12, a methyl group at C-13, a methyl group at C-14, and a methyl group at C-15. The stereochemistry is clearly defined with wedge and dash bonds.</p>
Escitalopram	 <p>The structure shown is a benzofuran derivative with a cyano group at the 4-position and a dimethylaminoethyl group at the 3-position. The stereochemistry is not clearly defined, and the dimethylaminoethyl group is drawn with a wavy bond.</p>	 <p>The correct structure of Escitalopram is a benzofuran derivative with a cyano group at the 4-position and a dimethylaminoethyl group at the 3-position. The stereochemistry is clearly defined with wedge and dash bonds.</p>
Esomeprazole	 <p>The structure shown is a benzimidazole derivative with a methoxy group at the 5-position and a dimethylaminoethyl group at the 2-position. The stereochemistry is not clearly defined, and the dimethylaminoethyl group is drawn with a wavy bond.</p>	 <p>The correct structure of Esomeprazole is a benzimidazole derivative with a methoxy group at the 5-position and a dimethylaminoethyl group at the 2-position. The stereochemistry is clearly defined with wedge and dash bonds.</p>
Pravastatin	 <p>The structure shown is a statin derivative with a hydroxyl group at the 3-position and a dimethylaminoethyl group at the 4-position. The stereochemistry is not clearly defined, and the dimethylaminoethyl group is drawn with a wavy bond.</p>	 <p>The correct structure of Pravastatin is a statin derivative with a hydroxyl group at the 3-position and a dimethylaminoethyl group at the 4-position. The stereochemistry is clearly defined with wedge and dash bonds.</p>

Table 2.6 Sample incorrect (left), and correct structures for several compounds retrieved as part of the initial structure resolution process.

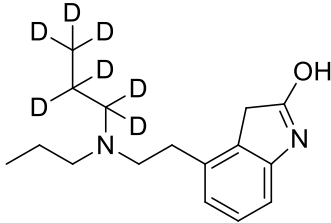
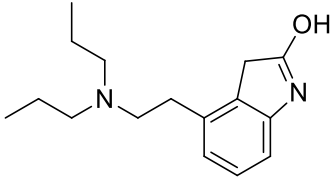
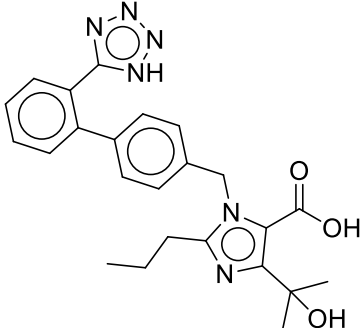
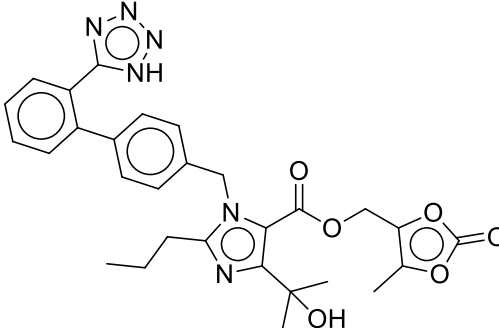
	Structure as Presented	Correct Structure
Ropinirole		
Olmesartan		

Table 2.7: Sample incorrect (left) and correct structures found in public structure sources.

## Discussion

### *Categorizing errors*

In the course of gathering these results, it became apparent that there are multiple categories of errors that arise in generating a curated chemical structure list from non-systematic identifiers. Foremost among these are problems inherent in the structures presented as correct on Internet sites. This does not mean that correct structures cannot be obtained, but it does underscore the need for caution in assuming that structures from any single source will be accurate.

In the structures examined, the most common inaccuracy was absent or incorrect stereochemistry (Table 2.6). In the case of escitalopram, one candidate structure reported no stereochemistry at all, in spite of the fact that escitalopram is explicitly the *S* enantiomer of citalopram. Likewise, there were cases where not all stereocenters were assigned, such as in

the case of azithromycin. The most complex problems were actually ones where there had been conflict in the literature over the correct assignment of configuration at stereocenters. For pravastatin, one team found early crystallographic papers that assigned a different absolute assignment of a chiral carbon on the cyclohexyl ring. It was assumed that the oldest crystallographic reports would be correct, and that subsequent divergence would be due to copyist errors. That was not the case.

One final source of problems in stereochemistry was assumptions made about the treatment of stereochemistry by software used to input, display, and/or transform the structures. A particular example of this is esomeprazole. Omeprazole is a proton-pump inhibitor that possesses a sulfinyl group connecting two ring systems. The sulfinyl sulfur forms bonds to three other atoms, but is not trigonal planar. It is tetrahedral because of the lone pair of electrons on the sulfur. Ordinarily the two enantiomers of a chiral sulfoxide interconvert rapidly at room temperature, but if the groups attached are sufficiently bulky, it may be possible to resolve the stereoisomers. In this case, the enantiomers were resolvable, and the pure *S* isomer was approved for sale as esomeprazole. We found that some chemical drawing programs either incorrectly assigned a trigonal planar geometry to the sulfinyl sulfur in esomeprazole or assumed that rapid thermal interconversion would always occur at that center. In one case, no stereochemical information was allowed to be drawn at the sulfur, and in another the stereochemical information was presented on the screen, but was stripped out of structures exported from the application in standard chemical formats.

Beyond stereochemistry there are, of course, other errors that can occur in name to structure mapping. These include structural deletions, isotopic substitutions, and incomplete product definitions. The first two situations did not occur in the final candidate lists used in the assembly of the gold list, presumably due to their relative rarity in the trusted sources used to construct the final candidate lists. They did occur in the public



databases that were subsequently compared to the gold list. Two examples of these are noted in Tablefigure 2.7.

A unique problem in this specific set of compounds that is potentially much more general is illustrated by the difficulties of assigning correct structures for premarin and premapro. The APIs in premarin and one of the two active ingredients of premapro are listed given simply as conjugated estrogens. Conjugated estrogens are sulfonated hormones extracted from the urine of pregnant horses. While the identities of some of the major components are known, a comprehensive list of all active small molecules present does not exist at the present time. The activity of the mixture is standardized by biological activity, rather than quantitative determination of specific compounds. Because the stoichiometry of the components is variable between individual batches, as is the degree and placement of sulfonation, a definitive statement on the structure of the API would, by definition, not be possible. Therefore these drugs were excluded from the study.

The question of tautomer assignment also can confound name to structure resolution. While the InChI contains support for tautomer identification, this is a newer feature. InChIs generated with older versions of the software and according to older standards may not have these features present. Even if they are enabled, it is not a panacea; the tautomer identification code in the InChI suite does not identify every possible tautomer. For example, both vardenafil and meloxicam can be drawn as multiple tautomers, and these tautomers do not have identical InChIs or InChI keys. When a primary tautomeric form has been identified in the literature, it is reasonable to expect that that form should be the one consistently used. However in the absence of a definitive determination, chemists will often make a prediction of the preferred tautomer based on empirical rules and past experience. There is a very real risk that different forms will be selected independently by two chemists, or that the experimentally correct tautomer will be normalized into another form by rules

attached to a chemical registry system. It is arguably best that compounds possessing the capacity for tautomerization be flagged in any comparison, and each tautomer enumerated (by either computational tools or human intervention) and separately checked.

One final category of problem to be considered is that introduced by the human element. The existence of *Chemical Abstracts*, Beilstein's *Handbuch*, and Gmelin's *Inorganic Chemistry* demonstrate that it is possible for chemical structure information to be accurately accrued, indexed, and disseminated without computer intervention, albeit at a very limited speed and with a correspondingly high labor cost. This, however, is not to say that the problems are trivial. The UNC workflow serves as a useful example of some issues involved in manual curation steps. Because of a desire to include information from sources that did not include machine-readable chemical structure information in them, such as the FDA Drug Label Archive and US and European Patent records, there were situations where visual comparison on structures, and manual transcription were inevitable. This led to several egregious errors in the proposed structures from the UNC team where it is presumed that the correct structure had, in fact, been located. This assumption was based on the very similar list of sources consulted by both the UNC team and the other teams, and on the inconsistency of structures within the team's results. For example, formoterol occurs in two places on the list, by itself and as a component of Symbicort. In the latter case, the UNC team provided a structure that was judged to be the same as the structure on the gold list. In the former case, however, an incorrect structure was given that was missing the central amino nitrogen. Taxol, on the other hand, only occurred once, but it also was missing a nitrogen atom although possessing no other errors. Finally, in ranitidine, an extra methyl group was added to a secondary amine to produce a tertiary amine that was not present in any of the databases searched for primary structure information. Since there was no primary source for this structure to be copied from, we assume that it is also an artifact of human inattention during structure transcription.

### Comparisons between workflows

The searches described here can be separated into two broad categories. The RSC and UNC workflows both were primarily manual and depended on a chemically-knowledgeable operator to make queries to various resources and compare the results of those searches. In contrast, the AZ and IMIM workflows were significantly more automated, with a set of names provided to a program that then performed the searches *en masse*

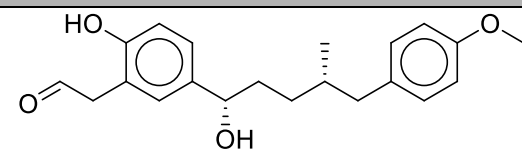
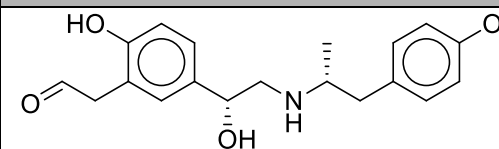
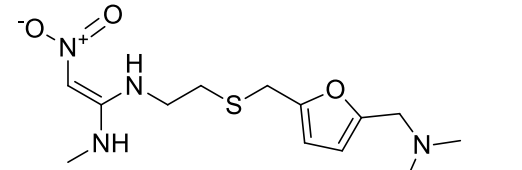
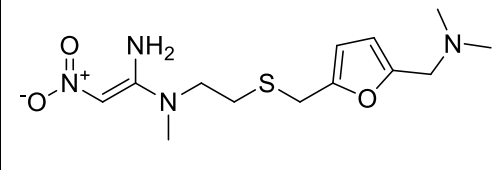
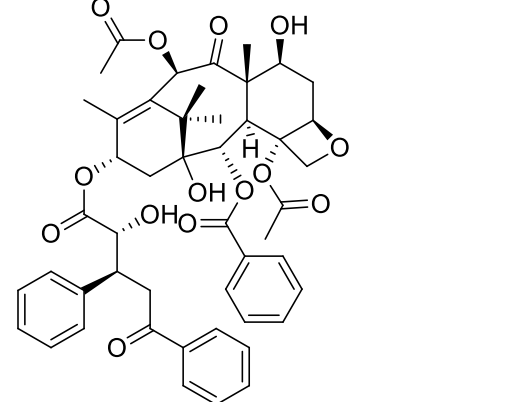
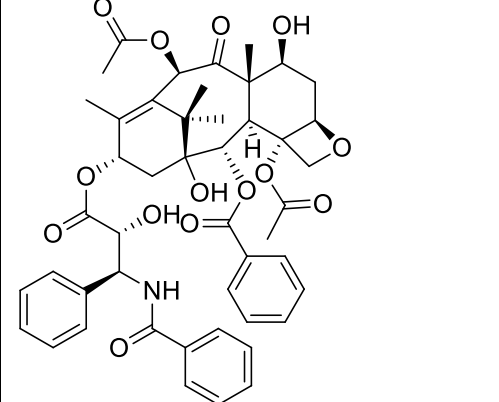
	Structure Presented	Correct Structure
Formoterol		
Ranitidine		
Paclitaxel		

Table 2.8: Errors (left) and corresponding correct structures (right) introduced by manual intervention in the curation workflow.

against either a proprietary database or a subset of sites found on the Internet. This division, at first glance, seems to be somewhat arbitrary and of limited utility. The “manual” workflows are first and third in overall accuracy of results, and the “automated” ones are second and fourth. However, the AZ workflow is like the “manual” workflows in one very significant way: the data automatically searched by software at AZ is already curated, both by the careful selection of proprietary and open sources of chemical structure and by the actions of other users at AstraZeneca, who are allowed and encouraged to correct inaccurate structures immediately. While the approach taken at IMIM incorporates source selection to some degree, the restriction to using only freely available sites appears to be too constraining for a fully automated search based on consensus structures to succeed. This is very possibly due to the relatively small number of original sources for chemical structure information that may be reused without restriction, and the equally limited number of chemically-literate Internet users who participate in large scale chemical curation activities.

This does not account for the difference in performance between the RSC and UNC workflows, which are both manual/human-centric. We account for the 4-fold increase in the number of total errors relative to the gold list by postulating that the increased use of visual comparisons and manual structure re-entry is responsible for this discrepancy. While it is not conclusive evidence for this hypothesis, the similar performance of the AZ workflow, which also minimized human intervention in the actual search, suggests that electronic structure handling and comparison help limit the introduction of new errors into an assembled data set.

It is also worth noting that all four teams agreed on the structure of a given compound 113 times as seen in Table 2.9. This represents a concordance rate of 75%. Conversely, there were two cases where only one team presented the structure that was ultimately determined to be most accurate (once by UNC and once by RSC).

<b>Number of groups in error</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>
<b>Percent Occurring</b>	74.8	16.6	7.3	1.3

Table 2.9: Relative frequency of numbers of groups reporting an incorrect structure for each search term

### *Standardizing electronic representations*

The InChI standard provides for an invariant, canonicalized string representation of molecular structure (compatible with ASCII and other standard 7 and 8 bit text-encoding systems) that addresses many problems of manual operations on chemical structure data. Unlike the SMILES molecular notation, each molecule is represented in a canonical form which is part of the original specification of the format. Stereochemistry and aromaticity are also represented in a standardized form. These features are intended to ensure that there is one, and only one, correct InChI representation for a given molecular species (unlike SMILES), allowing structural comparisons to be reduced to a simple string comparison. While the extended length of many full InChI representations of complex molecules can be problematic at first glance, this is not a major shortcoming as comparison of structures is a task most suitably delegated to software. When human comparison of structures is unavoidable, a hashed representation (the InChI key, which is uniformly 27 characters long) is also available. However, while InChI strings and keys are an improvement in chemical data handling, the current standard is not a panacea for all applications. As previously stated, the current standard does not necessarily represent all tautomeric forms of a structure in a uniform manner. It is also problematic that the InChI is much more opaque to manual interpretation and creation than older formats, such as SMILES. While it is true that the primary use of any computerized representation of a chemical structure is to allow

automated operations upon that structure, ultimately human end-users will need to interact with the resulting data when the processing is complete. It is relatively easy, even for non-specialists, to achieve basic competency in visually decoding a SMILES string to obtain some structural information about the molecule encoded. This is not a critical failure, but it may tend to impede adoption of the InChI standards in situations where there is extensive use of SMILES. In addition, there have been at least 3 “standard” sets of options for the creation of InChI strings that are to be hashed into an InChI key. While these options are primarily concerned with constructing canonical forms of resonance systems, and InChI keys constructed with non-standard options are possible, albeit tagged as non-standard, the need to consider these differences when comparing structures saved over an extended period of time does complicate matters. Finally, it has been demonstrated that there is at least one collision in the InChI key between two different structures. The two structures involved are both large and closely related, but the collision occurring at all raises the spectre of it occurring repeatedly. To fundamentally restructure the hash algorithm used would only make the standard versioning problem described worse. Again, this is not a reason to abandon the InChI, but certainly cause for developers and architects plan carefully before putting InChI-based applications into production.

In an era of electronic laboratory notebooks and the near-universal use of computers to prepare and edit technical manuscripts for the chemical biology literature, there is very little reason that a chemical structure should ever need to be transcribed between printed and electronic forms more than once. Even when hand-written laboratory notebooks are still utilized, entering structures into a manuscript is an essential part of the disclosure process, either for publication or for inclusion in a patent application. Similarly, since electronic catalogs of screening compounds are now the rule rather than a novelty, structures attached to high-throughput screening results are already available in machine readable formats. In the case of truly novel compounds, the chemist who planned and/or performed the actual

synthesis is more likely to know the presumptive structure of a new chemical entity than anyone else, and also to know the region of chemical space being explored which can help eliminate confusion about scaffold structure. This standard has become routine in bioinformatics, with most reputable journals requiring deposition of novel sequences in public repositories, and the development of the MIAME (Minimum Information About a Microarray Experiment) standard<sup>34</sup> that provides a structured format for information about gene expression studies. A similar standard, MIABE (Minimum Information About a Bioactive Entity)<sup>35</sup> has been proposed by a committee including representation from publishers, industry and academia, (notably with the input from the European Bioinformatics Institute, the maintainers of the ChEMBL database). While there has been institutional inertia resisting this standard, we believe that the adoption of the MIABE standard as part of the peer-reviewed literature publication process could improve the quality of public structural information by eliminating many occasions of manual re-entry of structures from the primary literature.

## **Conclusions**

Based on the experiences of four different teams utilizing their own preferred protocols, we conclude that it is possible to find accurate chemical structure information from publically accessible Internet sources, but not without careful consideration of the sources used. While these results indicate that there are several relatively accurate sources for chemical structure information available, none of them were universally correct, and none of them had all 151 structures being sought with no ambiguity or misannotation. Multiple independent sources will be necessary for the short to medium term to ensure that accurate structures are retrieved, especially when dealing with infrequently used compounds such as some of the specialized chemical biology probe compounds.

It is easy to point a disapproving finger at the compilers of chemical structure information and wag it disapprovingly for the errors found in databases. However, it is ultimately the end user who remains responsible for the accuracy of the structures used in their manuscript or models. Two of the three groups participating in this project not relying upon a previously curated database for structural information failed to achieve an error rate less than 5%. This implies that the correct structures are out there, but it's easy to miss them. More than anything else, using electronic representations of all molecules for transfers and comparisons is essential for the overall accuracy of a data set. There are times when visual comparisons and manual reentry of data are impossible to avoid, but these need to be held to a bare minimum and treated as rigorously as possible. At the very least, all manual comparisons should be double or triple checked, even going to the extent of making a scratch hard copy of the molecular structure so that atoms and bonds can be marked off as they are entered or checked.

While automated operations will eliminate errors introduced by human error, they only serve to accurately propagate errors when the data source being used is in error. This risk is hard to manage and almost impossible to eliminate. For compounds that are described in multiple, independent publications, cross-comparison of structures will provide a quick sanity check. For the vast majority of compounds, there will be only one published structure in the primary literature. Comparing the version extracted from a database to the original publication is as close to absolute verification as is possible without standards like MIABE becoming universal. Access to the original publication (or paying for access to SciFinder as a potential proxy) limits the general applicability of this approach. In instances where the data is only coming from one of a handful of sources, it may be useful to use a hard copy of the original paper and match electronic structures with those in the manuscript. Checking off each compound as its match is found is not foolproof, but it provides some confidence. If SciFinder access is available, it is possible to download the



structures of all compounds found in a given publication as SDF files and then to use those to match compounds, although this does entail some cost. As a compromise between effort, cost, and rigor, the best tradeoff for large datasets (over 100 compounds) may be to take a subset of the structures present and carefully compare them to the original publication. If more than one or two errors are found in the comparison, attempting to obtain the structures via an alternate method should be attempted.

The experience of the ChemoTargets/IMIM team (as graphically summarized in Figure 2.4) would tend to suggest that we have likewise not reached a point where automated queries of Internet resources or uncurated copies of those resources are capable of the same accuracy as searches guided by an experienced chemist. While the work of the ChemistryCentral project at AstraZeneca offers hope for this sort of lookup ultimately, manual or semi-automated curation is still an essential part of the search workflow, whether it occurs during the acquisition and initial construction of the database, or later in the lifecycle on a post-hoc and crowd-sourced manner.

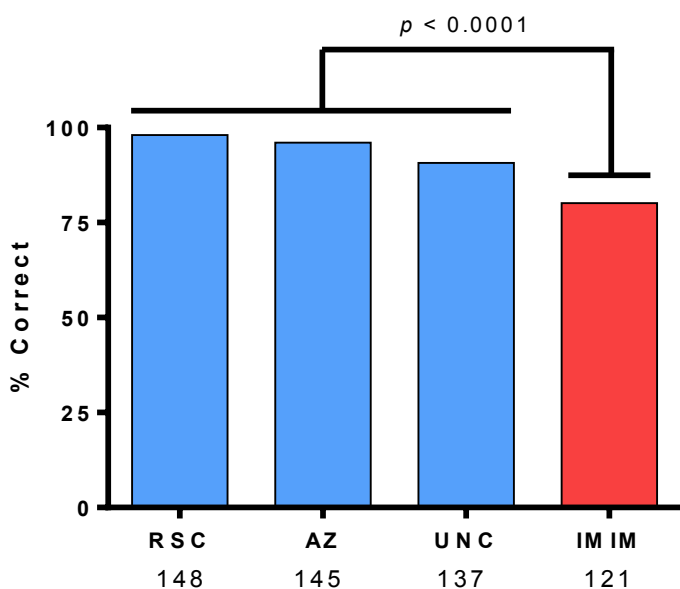


Figure 2.4: Summary comparison of accuracy of different structural curation workflows (raw data in table 2.4).

The manual reentry of structures from published documents into computer-readable form, and the use of visual comparison to verify the equivalence of two structures appear to be more effective in introducing errors into chemical structures than in helping to eliminate them. For example, the canonical representation of the structure of Taxol displayed in ChemSpider, and generated when the InChI representation is provided to ChemDraw and MarvinSketch are presented in figure 2.5. While all three of these have the same molecular weight and are topologically equivalent, it is a non-trivial process to visually compare them and verify that they are, in-fact, the same compound. Mentally rotating structures and determining which stereocenters are similar is a particularly complex process that offers ample opportunity for errors to be made.

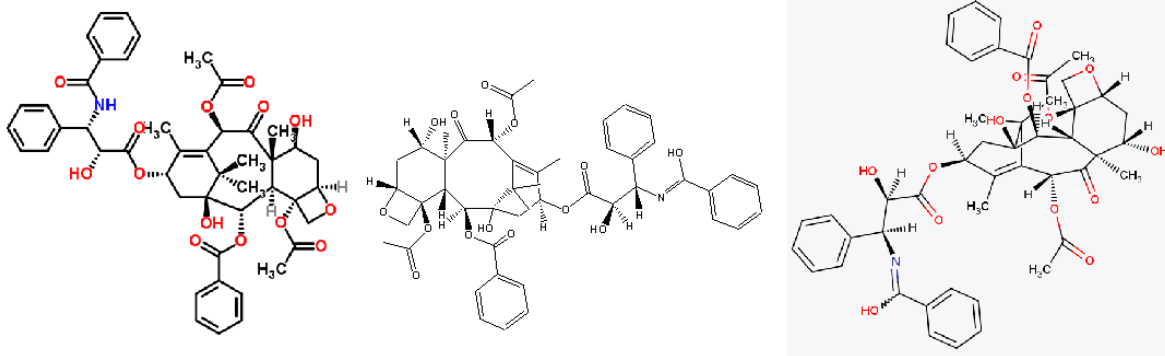


Figure 2.5: Chemical Structure of Taxol as displayed on ChemSpider (left) and as recreated with SMILES string obtained from Chemspider via ChemDraw (center) and MarvinSketch (right).

While we emphasize the role of best practices by end users who are assembling data sets for specific applications, these results are relevant to those who assemble large data sets containing chemical structure information for any purpose, even when the structural information is not central to the intended use of the data set. In particular, it is insufficient for a database to return the correct structure from a name query. Searches should also minimize (or better, eliminate) the number of incorrect and/or auxiliary and tangential answers that are

returned along with the correct one. While we insist on the importance of end-users carefully examining datasets for inconsistencies or errors prior to using them, the level of chemical and biological literacy among data scientists varies widely. There is no benefit in burdening the larger community with information that must be deconvoluted before use. Depending upon the nature of the data set, inactive components, excipients, and adjunctives should be either identified as such if relevant to the database's purpose, or eliminated if they are not germane to the data set. Furthermore, maintaining records of the source of structural data would be helpful when identifying cases where structures from a single source are collected from two different intermediate databases and then compared against each other as putatively independent verification of a single structure. Transferring structural information between open repositories is often not transparent and presents a challenge to finding truly independent confirmatory sources for structural validation. This problem is an area for careful future study, however, and not within the scope of this work.

## REFERENCES

- (1) Wiswesser, W. Historic Development of Chemical Notations. *Journal of Chemical Information and Modeling* **1985**, *25*, 258–263.
- (2) Brecher, J. Graphical Representation Standards for Chemical Structure Diagrams (IUPAC Recommendations 2008). *Pure and Applied Chemistry* **2008**, *80*, 277–410.
- (3) Norman, J. The Last Printed Edition of Beilstein is Published. <http://www.historyofinformation.com/expanded.php?id=1873> (accessed 24 Oct 2014).
- (4) CAS History. <https://www.cas.org/about-cas/cas-history> (accessed 24 Oct 2014).
- (5) Brown, A. C. On the Theory of Isomeric Compounds. *Journal of the Chemical Society* **1865**, *18*, 230–245.
- (6) Murray-Rust, P.; Rzepa, H. S.; Whitaker, B. J. The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World-Wide Web Information Exchange. *Journal of Chemical Information and Computer Science* **38**, 976–982.
- (7) Kaiser, J. Chemists Want NIH to Curtail Database. *Science* **2005**, *308*, 774.
- (8) Williams, A.; Tkachenko, V. The Royal Society of Chemistry and the Delivery of Chemistry Data Repositories for the Community. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 1023–1030.
- (9) Williams, A.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E.; Evelo, C.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: Semantic Interoperability for Drug Discovery. *Drug Discovery Today* **2012**, *17*, 1188–1198.
- (10) BioFocus Drug Discovery Databases Released to Public. <http://www.bio-itworld.com/biofocus-drug-discovery-databases-released-to-public.html> (accessed 24 Oct 2014).
- (11) Huang, R.; Southall, N.; Wang, Y.; Yasgar, A.; Shinn, P.; Jadhav, A.; Nguyen, D.-T.; Austin, C. P. The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing and Chemical Genomics. *Science Translational Medicine* **2011**, *3*, 80–96.
- (12) Williams, A. J. Reviewing Data Quality in the NCGC Pharmaceutical Collection Browser. <http://www.chemconnector.com/2011/04/28/reviewing-data-quality-in-the-ncgc-pharmaceutical-collection-browser> (accessed 24 Oct 2014)
- (13) Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today* **2011**, *16*, 747–750.
- (14) Halford, B. Bosutinib Buyer Beware. *Chemical and Engineering News*. **2012**, *90*, 34.
- (15) Elias, P.; Terrett, N. K. Pyrazopyrimidines for the Treatment of Impotence. Canadian Patent 2163446, July 02, 1998.

- (16) Supreme Court of Canada. *Teva Canada Ltd. v. Pfizer Canada Inc.*; [2012] 3 R.C.S.; 2012.
- (17) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50*, 1189–1204.
- (18) List of largest selling pharmaceutical products.  
[http://en.wikipedia.org/w/index.php?title=List\\_of\\_largest\\_selling\\_pharmaceutical\\_products&oldid=389022031](http://en.wikipedia.org/w/index.php?title=List_of_largest_selling_pharmaceutical_products&oldid=389022031) (Accessed 25 Oct 2014).
- (19) Dalby, A.; Nourse, J.; Hounshell, W.; Gushurst, A.; Grier, D.; Leland, B.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *Journal of Chemical Information and Modeling* **1992**, *32*, 244–255.
- (20) Wexler, P. The U.S. National Library of Medicine's Toxicology and Environmental Health Information Program. *Toxicology* **2009**, *198*, 161–168.
- (21) Wishart, D. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Research* **2006**, *34* (Supplement 1), D668–D672.
- (22) Wikipedia Foundation. Wikipedia. <http://en.wikipedia.org> (Accessed 24 Oct 2014)
- (23) Depoortere, H.; Zivkovic, B.; Lloyd, K. G.; Sanger, D. J.; Perrault, G.; Langer, S. Z.; Bartholini, G. Zolpidem, a Novel Nonbenzodiazepine Hypnotic. I. Neuropharmacological and Behavioral Effects. *The Journal of Pharmacology and Experimental Therapeutics* **1986**, *237*, 649–58.
- (24) Kuz'min, V.; Artemenko, A.; Muratov, E. Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. *Journal of Computer-Aided Molecular Design* **2008**, *22*, 403–421.
- (25) FDALabel - Full-Text Search of Drug Label Database.  
<http://www.fda.gov/ScienceResearch/BioinformaticsTools/ucm289739.htm> (Accessed 25 Oct 2014)
- (26) Brooksbank, C.; Cameron, G.; Thornton, J. The European Bioinformatics Institute's Data Resources: Towards Systems Biology. *Nucleic Acids Research* **2005**, *33*, D46–D53.
- (27) CAS presents "Common Chemistry". <http://commonchemistry.org> (Accessed 24 Oct 2014).
- (28) Gaulton, A.; Bellis, L.; Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research* **2012**, *40*, D1100–7.
- (29) O'Neil, M. J. *Merck Index*; 14th ed.; Merck & Co: Whitehouse Station, NJ, 2006.

- (30) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. Making Every SAR Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discovery Today* **2011**, *16*, 1019–1030.
- (31) Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research* **2004**, *32*, D277–D280.
- (32) Harmar, A.; Hills, R.; Rosser, E.; Jones, M.; Buneman, O.; Dunbar, D.; Greenhill, S.; Hale, V.; Sharman, J.; Bonner, T.; Catterall, W.; Davenport, A.; Delagrangé, P.; Dollery, C.; Foord, S.; Gutman, G.; Laudet, V.; Neubig, R.; Ohlstein, E.; Olsen, R.; Peters, J.; Pin, J.-P.; Ruffolo, R.; Searls, D.; Wright, M.; Spedding, M. IUPHAR-DB: The IUPHAR Database of G Protein-Coupled Receptors and Ion Channels. *Nucleic Acids Research* **2009**, *37*, D680–D685.
- (33) InChI Software Downloads. <http://www.inchi-trust.org/downloads/> (Accessed 25 Oct 2014).
- (34) Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C.; Causton, H.; Gaasterland, T.; Glenisson, P.; Holstege, F.; Kim, I.; Markowitz, V.; Matese, J.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. Minimum Information about a Microarray Experiment (MIAME)—toward Standards for Microarray Data. *Nature Genetics* **2001**, *29*, 365–371.
- (35) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; Glen, R.; Grigorov, M.; Hammond-Kosack, K.; Harland, L.; Hopkins, A.; Larminie, C.; Lynch, N.; Mann, R.; Murray-Rust, P.; Piparo, E.; Southan, C.; Steinbeck, C.; Wishart, D.; Hermjakob, H.; Overington, J.; Thornton, J. Minimum Information about a Bioactive Entity (MIABE). *Nature reviews. Drug Discovery* **2011**, *28*, 661-669.

## Chapter 3: Biological Data Curation

*“Hell is other people(’s data)”*

With apologies to Jean-Paul Sartre

### **Summary**

It has been demonstrated that the quality of data used as the basis of a QSAR model is a primary factor controlling the accuracy and external predictivity of that model. While the chemical structures used in a QSAR model can be objectively judged as correct or incorrect, biological activities are often measured as numerical values that arise from the interaction of multiple factors and that are gross macroscopic reflections of a microscopic statistical ensembles. As an example, binding affinities may depend on the expression system used for a given receptor, the assay technology used, the reference ligand used for competition (if any), and the skill of the scientist performing the assay. A cited rule of thumb in pharmacology suggests that if two different laboratories measure the binding affinity for a given ligand-target pair and their results are within one order of magnitude of each other, then the two results are considered “identical” or, at least, replicated. Because binding affinities are among the most frequently modeled activities in drug discovery, we have mined the medicinal chemistry literature and sought to determine the distribution of binding affinity values and the uncertainty associated with them. Thus, we have determined how many different compounds have been assayed for a given target, how often any given ligand has been assayed, and how many affinities

are reported within a data set from any given organization. These distributions quantify the necessary frequency of combining of data from multiple sources to derive a single QSAR model.

Because binding affinities are one of the most commonly used types of activities in drug discovery, and  $K_i$  values are the most comparable measure of affinity, the first part of this study was to select a subset of data from ChEMBL where the data only included  $K_i$  values measured for a single small molecule at a well-defined biological target. By using only ChEMBL data that was been abstracted from the peer-reviewed literature, we eliminate results derived from HTS data sets that generally have lower reproducibility and would tend to skew the results towards a particular set of molecular targets. Once this subset was created, entries that exactly duplicated another entry in the data set (having the same ligand, target, activity type, activity units and activity value) were removed in order to eliminate data that was most likely copied verbatim from an earlier source. This data was then summarized to determine how many times any particular molecule was found in the literature, how many distinct molecules appeared in each peer-reviewed publication, and how many different ligands had been assayed against each target. Once these data were assembled, the activity data was cross-joined with itself such that all possible pairs of activities for each ligand and target were paired. This table was then converted to use standard  $pK_i$  values for affinities and the absolute difference between each pair of compounds was taken. All differences less than 0.05 log units were considered to be probable copies of data in different unit systems and were removed from the data set. All differences greater than 12 log units were dropped as being too large to be anything other than errors in abstracting and editing. The remaining values were plotted as an empirical cumulative distribution function, and this distribution was examined to determine the mean, median, quartile, and 80% range values for the distribution. A non-random subset of pairs showing over 3 log units difference was examined to identify common sources of large differences.



On average, each compound reported in a binding assay in the medicinal chemistry literature appears 1.17 times. Less than 10% of all compounds reported have multiple binding affinities reported for any target. Similarly, the average paper reporting binding affinities in the medicinal chemistry literature will have 18.8 affinities for different ligands reported in it. The average target will have a mean of 179.8 ligands assayed for  $K_i$  against it, but 50% of all targets have fewer than 26 ligands assayed, yielding a highly skewed distribution. This leads to the conclusion that independent replicate assays of binding affinities are rare enough in the literature that comparing such measurements will not be sufficient to detect errors in binding affinity values. Also, with only 20 compounds or less reported in a single paper on average, it is important to examine robust methods for combining data from multiple sources when developing predictive QSAR models will typically require at least 80-100 compounds. Finally, opportunity for further investigation exists when almost half of all targets have only a single (at most two) paper reporting their binding affinity data. By examining the empirical cumulative distribution function, we estimate that the mean uncertainty associated with a binding affinity values is 0.48 pK units, with a median of 0.29 units. This result compares favorably with that completed recently by researchers at Novartis who estimated the error to be 0.44 pK units, with a median of 0.34 units. This curve also suggests that 87.4% of all reported binding affinities lie within 1 pK unit of a hypothetical “true” value. While this finding validates the 1 log unit rule of thumb for comparing affinities obtained in different laboratories or by different methods, it also implies that attempting to predict binding affinities to less than a log unit of precision when data from multiple sources are used will be problematic. A workflow is proposed for further testing that attempts to minimize the uncertainty in differences between distinct binding affinities and generate training sets for QSAR modeling which show increased accuracy in predictions.

This work provides useful information not appearing elsewhere about the frequency of assaying of ligand-target pairs, the number of binding affinities reported at a single target at once, and the size of training sets available for modeling ligand-target pairs in the primary

literature. The estimated uncertainty in a binding affinity measurement confirms a value recently reported in the literature that used an earlier version of ChEMBL and a more labor-intensive process and was based on significantly fewer data points. By generating the binding affinity pairs in place as part of the data extraction, larger datasets can be analyzed without manual intervention. Finally, previous work on data curation specifies a need for removing duplicate values for given target-ligand affinities, but it does not specify an algorithmic workflow to accomplish this task. This chapter proposes a rational workflow to accomplish the deduplication. While its ultimate utility remains to be determined, no other workflow has been formally proposed to accomplish this objective.

## **Introduction**

When considering the accuracy of the depiction of a molecular structure, there may be ambiguity arising from the specific conventions of the representation method used. However, most chemists would agree that there is one absolute, Platonic structure for any given molecule (or at least an ensemble of a few structures representing the extreme values of interconverting forms). That is, there is a correct molecular structure that we can know and accurately reproduce to an arbitrary precision. When we turn our attention to the binding affinity between small molecules and biological macromolecules, which is a central measure in many facets of drug discovery and chemical biology, things become murkier.

When a typical small molecule binds to a protein target, there is a change of free energy of binding ( $\Delta G_{\text{binding}}$ ) of somewhere between 1 and 10 kilocalories/mole, or on a per-event basis, about 2 attojoules ( $10^{-18}$  J). With the most sensitive thermocouples available capable of detecting temperature differences of one hundred billionth ( $10^{-9}$ ) of a degree, a 4 microlitre calorimetry cell charged with protein and a 100nm ligand to that would have to have approximately 240 billion binding events occur to register any net heat evolution. This is an extremely large number of events, which we cannot hope to be measured discretely. Statistical treatments are essential.

When issues arising from differing expression systems, cell lines, or even extraction protocols are included on top of this inherent variability for nominal biological replicates, there is a surfeit of uncertainty involved in estimating values of free energy of binding on a molecular scale. Even when new technologies promise to eventually detect binding events in a sample containing a few thousand protein molecules on a microsecond timescale, the measurement of binding affinities and rates is inherently a statistical process. This is not to say that binding affinity studies are inherently inaccurate. A well-planned protocol with properly instrumentation and a consistent protein expression system will allow highly reproducible binding affinity data that does not exhibit significant drift over time. These data can accurately represent the relative affinities of different ligands at the receptor or the relative affinity for a given ligand at a family of protein targets, within a finite level of precision. Nevertheless, anyone seeking to define **the** absolutely correct and final inhibition constant for any pairing of ligand and target is on a fool's errand.

A commonly cited rule of thumb among pharmacologists is that, when considering multiple independent binding assays of a given ligand-target pair, the results should be considered identical if the two affinities differ by less than 10% (when both assays were performed in the same laboratory) or by less than an order of magnitude (if the experiments were performed in separate laboratories)<sup>1</sup>. While this may seem to be a very lax metric, it actually reflects the difficulty of measuring the relative concentration of a compound in two samples at low concentration. In general form, an assay will consist of allowing a given quantity of a ligand to equilibrate with a fixed amount of the target, and then separating the target from the supernatant liquid and determining how much of the ligand remains unbound, or with a radioisotope-tagged ligand, measuring the amount of competent ligand remaining with the target.

In principle this is easy, but the reduction to practice quickly becomes complex, especially in high-affinity systems where the  $IC_{50}$  is under 10 nM. One traditional method for a protein-ligand binding assay involves indirectly measuring the amount of a reference ligand of known affinity displaced by successive amounts of the test ligand. By measuring the radioactivity of the reaction solution (after the protein target is removed), it is possible to measure how much of the reference ligand remained bound and, by extension, the ratio of binding affinities of the two ligands and ultimately the binding affinity for the ligand of interest<sup>1</sup>. While great advances in pharmacology and medicine have been accomplished using these techniques, it represents an extended chain of values with multiple sources of error propagating through the process (measurements of concentration and radioactivity, assuming all hot ligand is either free in supernatant or bound stoichiometrically, accuracy of binding affinity of hot test ligand). Newer techniques, such as isothermal calorimetry and SPR, measure the binding event without a need for radioisotopes and compatible test ligands, but these assays are not inherently amenable to high-throughput development and still have uncertainties of their own. No matter which technology is used, the fundamental difficulty with measuring binding affinities remains the same: measuring the  $IC_{50}$  of a ligand to within half a log unit requires detecting changes in ligand concentration of less than 25%; for a quarter log unit, changes of 14% from the actual  $IC_{50}$  value must be measured<sup>1</sup>. Ultimately, the measurement of ligand binding affinities is highly sensitive to individual laboratory technique, the protocol employed, and the choice of reference compound or direct detection method.

When building a QSAR binding affinity model, a modeler usually is constrained to use what data is already available, as “beggars” are rarely in a position to be particular. While it may be possible to obtain additional data, either by virtue of a particularly close collaboration or by working in a research group with both experimental and computational members, repeating already published studies or generating significant amounts of new binding affinity data (especially when the ligands of interest are not commercially available) will not be possible for

the most part. Obviously, it is desirable to use reliable data and set aside that which is considered suspect. However, without affinity values from more than a single source, is it possible to identify problematic data by inspection of a single research group's assay values? At the same time, there are cases where more than one affinity has been reported for a given ligand-target pair. This may seem to be an improvement over the former state of affairs, but it actually just raises new questions: "Which of the duplicated values should be chosen for use in a model?" and "How much of a difference in affinity values is allowable before that decision actually has an impact on model performance?"

A classical approach to make data consistent from multiple labs would be to construct a calibration curve for results of the same ligand target pairs. However, it is uncertain whether there will be enough reliable replicates from different laboratories to actually use this method. At a minimum, affinities for two compounds against the same target determined by different providers would be enough data to start with, but the correction would be limited to a straight line and would probably perform inconsistently over the usual range of affinity values. Ideally, five or ten distinct compounds would be assayed at the same target by two providers and distributed affinities over the range of possible values ( $pK_i$  values between five and eight or nine under most circumstances).

These questions are best answered by a large-scale analysis of published affinity values, preferably including both HTS libraries and traditional SAR series in addition to one-off reports of a single ligand's affinity for a target of interest. However, it would be impractical to manually scan the major medicinal chemistry journals and transcribe information from each paper into a machine-readable format. Even an exceedingly small random sample could take months or years to generate. Fortunately, a ready source of such data is available in ChEMBL. ChEMBL is a useful proxy for the medicinal chemistry as a whole, as its primary data sources are the four journals in which SAR datasets from academic and industrial laboratories are most likely to be

published. It also contains DrugMatrix high-throughput screening (HTS) results, HTS values for the affinity of common ligand target pairs. The latter is significant as large HTS campaigns generally attempt to cover a broad swath of chemical space by use of focused or screening libraries. The resulting data are inherently sparse with few cases of analog series based on a common scaffold with extensive variation in substituent groups. Individual SAR series, on the other hand, are denser series of compounds that were chosen to attempt to approach a maximal binding affinity for a given region of chemical space against a given target. Because of the proximity of these compounds to each other in chemical space, it is more readily possible to observe activity cliffs in a SAR series. Activity cliffs are ordinarily considered problematic as they represent a discontinuity in the response function that makes accurate modeling in its vicinity difficult.<sup>2,3</sup> However, activity cliffs also represent opportunities for drug design as the presence of such a cliff may indicate a change in ligand binding mode at the target (suggesting the opportunity for a scaffold hop to novel compounds) or that a local maximum binding affinity is near (because a small change to the ligand has made it unable to fit within its binding site).

Since its debut in 2008, ChEMBL<sup>4</sup> has become a major resource for researchers, not only those interested in particular target or phenomenon, but also for those interested in larger scale patterns and trends in known chemical space. In 2012, a group at Novartis Institutes for Biomedical Research published a study of their analysis of errors in ChEMBL 12<sup>5</sup>. Their primary goal was to estimate the best correlation constant possible for regression QSAR models built from  $K_i$  data in ChEMBL when using data collected independently by two or more research groups. To this end, they collected all the biological replicate experiments for binding affinity present in ChEMBL 12, cross-referenced them to PubMed to identify measurements originating in the same group and eliminate those duplicates, and applied a series of filters to eliminate measurements that they considered likely to have significant error. The remaining activities were then grouped with their biological replicates and statistically analyzed. They calculated a mean experimental uncertainty for binding affinities in ChEMBL as being 0.44 pK<sub>i</sub> units with a

median of 0.34 pK<sub>i</sub> units, with a maximal correlation coefficient of 0.81 observed over their highly curated data set.

Even with this extensive analysis, the Novartis paper is not a definitive statement of the uncertainty for published small molecule binding affinities. In their determination to offer a highly accurate estimate, the authors went to great lengths to eliminate duplicate reports of a single experimental value. While this is a reasonable assumption to make, they also excluded reported values which were offset by exactly 3 or 6 orders magnitude, reports appearing in reviews, results which were not already labeled as K<sub>i</sub> measurements, and results where K<sub>i</sub> was a negative value. While all of these are reasonable filtering steps, this process also has a strong tendency to eliminate outliers in the literature, and such values might or might not be identified by those assembling data sets for molecular models. Further, it seems likely that their reported values underestimate the effective uncertainty of a single binding affinity reported in ChEMBL. By eliminating all values that are not already in K<sub>i</sub> form, they also excluded a significant number of results in ChEMBL that were already reported in pK<sub>i</sub> (or Log K<sub>i</sub>) format, artificially limiting the size of the pool of biological replicates.

The Tiikainen paper of 2012 has been discussed elsewhere in more detail (see Chapter 1). It is sufficient to state here that this work was primarily concerned with those errors in compiled databases that are introduced by the compilation process, and this particular error rate is somewhere between 5 and 8% of all structures and values reported in one of three large biochemical activity databases<sup>6</sup>. This finding is both bad and good news – bad in that it is previously known that 5% error is enough to cause models to become non-predictive<sup>7</sup>, but good in that the problem is arguably amenable to a curation process that can reduce the error rate by 10%, 20%, or more.

In light of these investigations, the question remains as to whether it is possible to design a workflow that can consistently minimize the errors in a bioactivity data set and increase the

likelihood of predictive models being generated. This work, therefore, was not conceived to attempt to generate a more accurate value of the uncertainty inherent in ChEMBL binding affinity data or to more completely map out all the errors in that subset of the biomedical research literature. Our primary goal in this work is to understand the particular pitfalls of extracting biological activities from the primary literature and to propose a workflow which helps reduce the errors inherently present in such data to a more manageable level when used as part of a QSAR modeling workflow. As such a workflow is targeted at less-experienced cheminformaticians and modelers, we intend to take a more naïve approach to data assembly and curation, hoping to capture a more typical cross section of issues that would be faced by the intended end-users of such a process. We will also confine ourselves to the using data from ChEMBL, as it is open access and cost-free, and more likely to be the first stop to find data by part-time modelers than a commercial database.

## **Methods**

### *ChEMBL Logical Organization*

ChEMBL is provided by the European Bioinformatics Institute of the European Molecular Biology Laboratory as a set of tables for installation in a Relational Data Base Management System (RDBMS). In a RDBMS, different parts of the overall data are stored in smaller tables that are cross-referenced by unique identifiers attached to each row of each table. Complex reports are created by combining data from different tables into a single file under conditions and constraints.

The primary goal of ChEMBL is to consolidate reports of biomolecular activity into a freely-available, machine-searchable repository. While its primary focus is on reports arising in the primary research literature, it also mirrors the DrugMatrix Toxigenomics database (originally developed by Entelos, currently supplied by NIEHS) and confirmatory screen data



from the PubChem project, and hosts several large screening datasets donated by the pharmaceutical industry, most related to malaria, tuberculosis, and other diseases of the developing world.

In the case of ChEMBL 14, there are 29 individual tables. These tables contain multiple types of information, including the version of the database in use, how far different compounds have advanced in clinical trials, hierarchical tables describing different salts prepared of organic acids and bases, cross-references for different entities' in house code names, and the details of what organisms were used in a particular assay. This information is cross-linked via a set of core data tables present in ChEMBL corresponding to small molecule ligands, targets/sites of action, assays, documents, and activity values. In ChEMBL14, there is information relating to 1,213,242 compounds, 9,003 targets, 644,734 assays, 46,133 documents, and 10,129,259 individual bioactivities. Given that there are 10,922,817,736 possible combinations of targets and compounds, the data can be considered sparse with only 9.3% of possible combinations represented.

The primary table needed to identify and compare biological replicates and the relationships between them are shown graphically in figure 3.1. More specifically, the chembl\_14.activities, chembl\_14.docs, chembl\_14.assays, chembl\_14.compound\_records, and chembl\_14.target\_dictionary tables have human-readable details of the data abstracted from the relevant literature. Because of the size of the target\_dictionary table, which includes protein sequence information, it is not directly cross-referenced by the assays table, but instead a junction table, assay2target is used. This table cross-references targets with assays which quantify activities at those targets, while limiting the amount of data that must be pulled in from either table, especially when information about a small subset of assays or a small number of targets must be retrieved. In this case, the assay2target table is particularly important, because it also contains two other fields which are used to describe what evidence there is for a given

target to be the actual target of a given assay and what kind of relationship exists between the known and putative targets. These two fields are named “relationship\_type” and “confidence score.” They contain only a single character each because they both reference external tables which define the meanings of those shorthand codes. The codes are defined in tables 3.1 and 3.2. While their meanings overlap, the key distinction is that the confidence score represents how certain the assignment of a target for a ligand is, while the relationship type indicates the quality of the evidence used to assign that target in ChEMBL.

<b>confidence_score</b>	<b>Definition</b>	<b>Localization</b>
0	Target unknown or has yet to be assigned (default)	Unassigned
1	Target assigned is non-molecular	Non-molecular
2	Target assigned is subcellular fraction	Subcellular fraction
3	Target assigned is molecular non-protein target	Molecular (non-protein)
4	Multiple homologous protein targets may be assigned	Multiple homologous proteins
5	Multiple direct protein targets may be assigned	Multiple proteins
6	Homologous protein complex subunits assigned	Homologous protein complex
7	Direct protein complex subunits assigned	Protein complex
8	Homologous single protein target assigned	Homologous protein
9	Direct single protein target assigned	Protein

Table 3.1: ChEMBL 14 Confidence Scores for target assignment quality assessment

<b>Relationship Type</b>	<b>Meaning of Code</b>
<b>D</b>	Direct protein target assigned
<b>H</b>	Homologous protein target assigned
<b>M</b>	Molecular target other than protein assigned
<b>N</b>	Non-molecular target assigned
<b>S</b>	Subcellular target assigned
<b>U</b>	Target has yet to be curated (default)

Table 3.2: ChEMBL Relationship Types

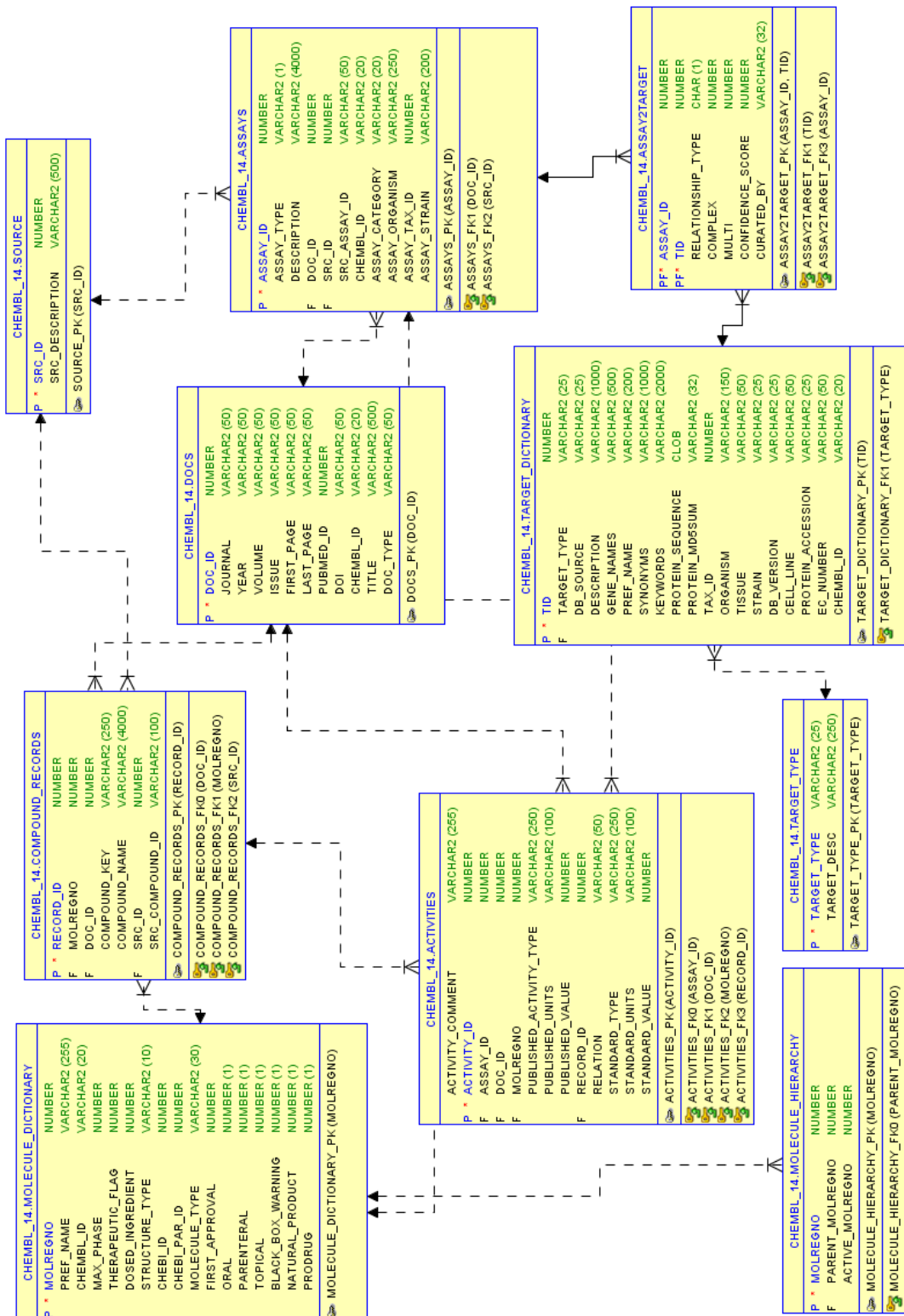


Figure 3.1: ERD for core tables of ChEMBL 14, with all columns and constraints shown

### *Initial Data Assembly*

Data for this analysis was taken from a local copy of ChEMBL 14, obtained from the European Bioinformatics Institute and installed in a local instance of Oracle 11. Data points were generated from an SQL script (Appendix 3) and exported to a tab-separated value formatted file. An initial table combining target, ligand, document, assay, and activity identifiers was built with constraints imposed in order to limit the presence of less reliable data and restrict the comparisons to comparable experiments. In particular the following conditions were imposed:

- The relationship type was required to be “D”, “H”, or “M”
- The target confidence score was required to be 3 or more
- The publication type was required to be “PERIODICAL”

This table was then self-joined to create a raw list of activity pairs by imposing additional join constraints on each resultant line (here consisting of lines *a* and *b* from the source table)

- The target ID for *a* and *b* must be equal
- The ligand ID for *a* and *b* must be equal
- The document ID for *a* and *b* must not be equal
- The activity ID for *a* must be less than the activity ID for *b*

An additional loose constraint that the standard activity type was required to contain the substring: “KI” was applied. This was not meant to instantly isolate all binding constant identifications, but rather to make a simple first cut to eliminate as many non-binding constant measurements as possible. The resulting output of this cross-join was then saved to a tab-delimited format and manually curated in Excel to remove inappropriate entries and provide an estimate of the quantity and types of issues present.

Published units and standard assay types were examined for each entry. During an initial pass, entries referencing a percentage inhibition, weight-based dosages, or other intractable/incompatible units were removed, as were entries with missing values, non-numeric results, or floor or ceiling values. Subsequently, all values were converted to pK<sub>i</sub> measurements based on molar concentration units. Entries where one or both paired values were not readily converted by automated rule, *e.g.*, pK<sub>i</sub> values with units attached to them, negative K<sub>i</sub> or pK<sub>i</sub> values, and physically unreasonable K<sub>i</sub> values, as well as a small sample of the successfully converted values were identified and subjected to manual confirmation based on the contents of the original literature report. Once those assessments had been completed, either by manual recalculation of the proper K<sub>i</sub> values or by deletion of inappropriately duplicated values, the difference between the two pK<sub>i</sub> values on each line was calculated. Each entry where the absolute difference between the two values was less than 0.05 pK<sub>i</sub> units was removed from consideration, as it was presumed that any difference that small was due to either direct copying of an earlier value or reuse of an earlier value with alteration in units or precision.

This approach was useful for identifying gross problems with comparing individual affinity measurements, and it achieved the goal of casting as wide a net as possible over the totality of the indexed literature. It became clear, however, that this methodology was insufficient for more detailed estimates of error. Because duplicated values were not removed before being used to compute activity pair differences, it over-valued the effect of repeatedly cited values and inflated the apparent variability between different determinations.

#### *Refinement of Data Assembly*

A second query was devised in an attempt to eliminate this over-counting of duplicate entries. We deemed it essential to retain the automated pair generation process, but duplicate values needed to be removed prior to this process. Furthermore, we decided that the simplest way to accomplish these goals was to separate the extraction of raw data from ChEMBL from the

deletion of the re-cited data and the actual construction of the activity pair differences. A three-part process was developed and employed to perform this extraction. In the first phase, the unique internal identifiers for targets, assays, activities, ligand molecules, and documents for the entirety of ChEMBL were joined into a single table along with the standard units and values for each entry, and filtered by the same single entry criteria used in the first process (target confidence score, target relation type, and publication type). In addition, a stricter filter was applied to the standard assay type for each record, removing all records that were not identified as  $K_i$ ,  $pK_i$ ,  $\text{Log}(K_i)$  or  $-\text{Log}(K_i)$  values. The results of this query were saved as a scratch table. A second query was then used to identify entries with the same target and molecule identifiers, and the same standard values, units, and assay types, but different activity identifiers (indicating that each value was formally a different entry). For each of these groups, the lowest activity identifier was retained into a second scratch table, eliminating entries where an affinity value was exactly the same as an earlier entry. While this did not eliminate all duplicates (such as those where a conversion between  $K_i$  and  $pK_i$  or a complex change of units had occurred), it did minimize the presence of duplicates before the primary cross-join.

This second temporary table was then joined to itself with the same conditions as the self-join in the first instance (same target and ligand, different document identifiers, and first record of each joined record having a smaller activity identifier than the second). At this point, addition information was added to the joined lines to make it more intelligible to human eyes (target and ligand names, species information, activity comments, etc). This final table was saved as a tab-delimited text file and exported for manual curation in Excel (version 2010, Microsoft Corporation, Redmond WA).

In Excel, canonical  $pK_i$  values were calculated for both molecules in each row (or record). When a syntactically correct  $pK_i$  record already existed (positive valued, with either a null or M entry for standard units, and some variant of  $\text{log}(K_i)$  or  $pK_i$  in the standard type field), that value

was copied verbatim into the canonical  $pK_i$  column (or field). Where possible, automated rules were used to convert groups of rows that shared common standard types and units to  $pK_i$  values in a consistent manner. In addition, records with incompatible units or requiring more extensive effort to convert (such as weight-based measures and percentage values) were removed; records which had physically impossible measurements (such as negative concentrations) were either corrected or removed after consultation with the primary publication. Records with an assay\_type of  $\text{Log}(K_i)$  or  $-\text{Log}(K_i)$  and an inconsistent value were presumed to be sign errors and multiplied by -1. After well-behaved  $pK_i$  values for all entries were calculated, the absolute value of the difference of between each pair of values was calculated ( $\Delta pK_i$ ) and added as a new column. The resultant spreadsheet was then sorted on the  $\Delta pK_i$  in increasing order, with all values of  $\Delta pK_i$  of 0.05 or less being removed as probable duplicate values. The final results were then saved as a tab-delimited file and imported into R<sup>8</sup> (version 3.0.2) for further analysis.

Information about the number of distinct compounds tested against individual targets, the number of different publications in which an individual compound appeared, and the number of distinct molecules present in each individual publication was extracted either from the final paired-affinity difference data or directly from ChEMBL. For values extracted directly via SQL, values were obtained for both a subset of the ChEMBL data mirroring the data in the paired affinity difference values derived from applying conditions on the search which mirrored the constraints applied in the first search of the paired affinity difference generation process (target confidence score, target relationship, and “B” as the assay type), and for ChEMBL as a whole with no constraints. For the limited subset, the output was structured to return all unique pairings of the two indices of interest and then condensed into aggregate forms in R. Owing to the overall size of ChEMBL, the data were consolidated into count data in SQL before export to R.



## Results

### *Distribution of magnitude of differences in paired binding affinity measurements*

After manual curation, the distribution of the calculated differences can be plotted with a kernel density plot, as shown in figure 3.2. While the distribution is roughly exponential in form, there are also large-scale irregularities that deviate from this distribution. In particular, there is a slight broadening of the distribution around 1 pK unit and a series of more pronounced shoulders around 2 pK units. Finally, while the overall distribution appears to have effectively returned to zero by 4 pK units, there are three notably visible peaks to the right, at the integral values of 3, 6, and 9 pK units.

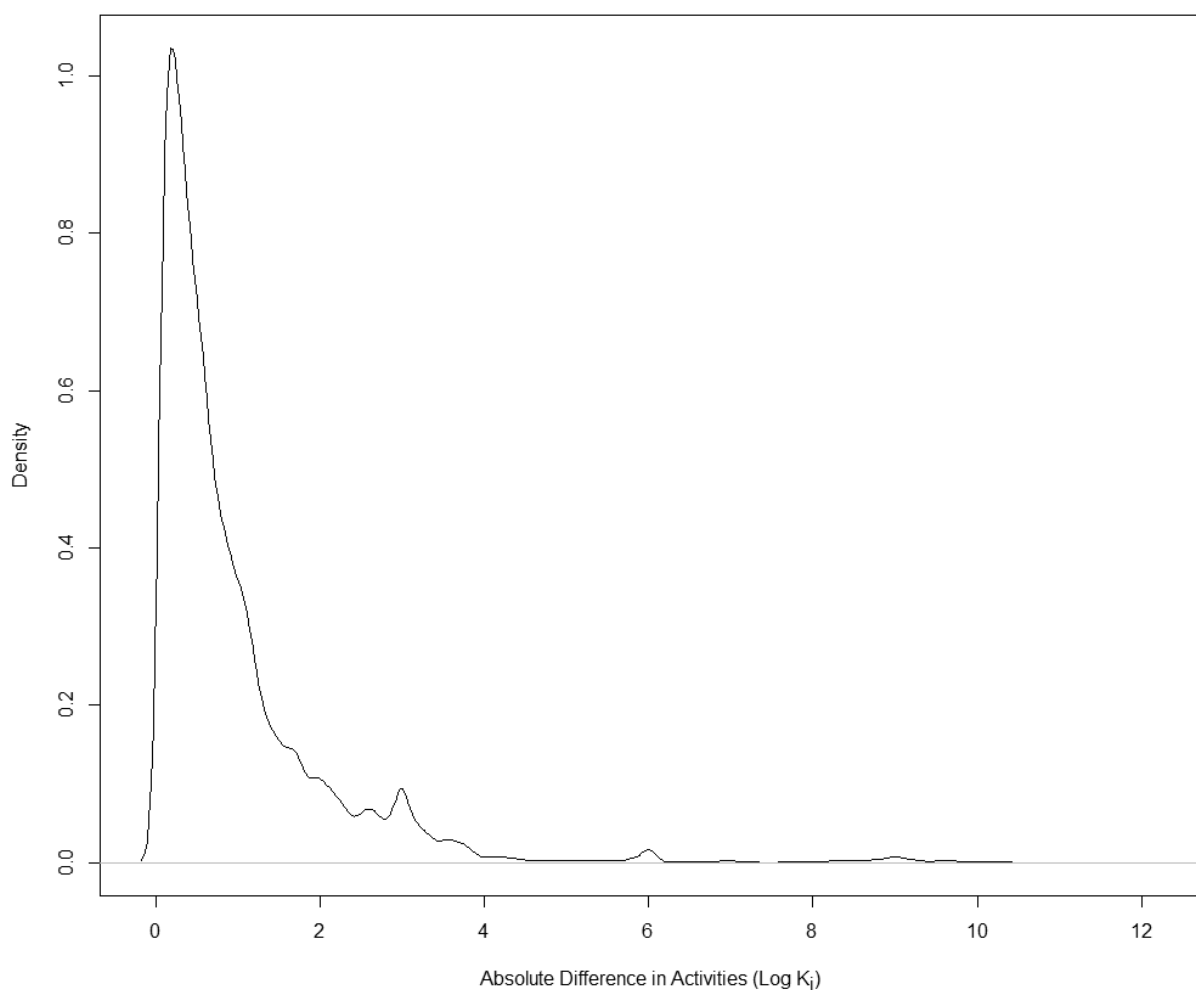


Figure 3.2: Distribution of absolute activity differences for biological replicate pairs identified in ChEMBL 14

### *Distribution of binding affinity data*

As stated previously, ChEMBL 14 contains 1,213,242 distinct compounds (1,384,479 records in the compound library), 10,129,256 bioactivities, 9,003 targets and 46,133 different documents. A naïve calculation would suggest that there ought to be 135 different molecules assayed against every target and that each document ought to contain 3 different molecules. However, ChEMBL contains many different types of assays from the medicinal chemistry literature and that for some drug classes, most notably antibiotics and antineoplastics, cell-based assays are much more common than biochemical protein binding assays. This raises the question of whether the binding affinity data in ChEMBL is distributed throughout the literature similarly to the entirety of ChEMBL. It also is a reasonable and arguably more important question whether the data in our selected subset of the binding affinity data (binding affinities independently replicated in the primary literature) is distributed similarly to the whole of binding affinity data.

We can characterize the coverage of the paired subset of the binding affinity data relative to all ChEMBL binding affinity data in terms of three scalar metrics: the number of distinct ligands, targets, and documents present in the subset. These are supplemented by three distributions: the ratios of ligands to targets, ligands to distinct documents, and documents to distinct ligands. These distributions describe different aspects of the data which are relevant to the QSAR modeling process. The number of distinct ligands for each target describes the maximum number of data points available for creating a model of a given target and serves as a proxy measure of whether a QSAR model of a target is feasible. The number of ligands per distinct document illustrates the dispersion of data throughout the literature. As will be discussed later, having all the affinity data for different compounds in distinct publications limits the number of replicate affinity measurements (which allow direct comparison of different protocols and/or different research teams). The number of

documents for a distinct ligand is essentially the maximum number of replicate values that would be available for binding affinities of a given ligand, and therefore the maximum number of different assay protocols that might be correlated by this data point.

	<b>Paired Affinities</b>	<b>Binding Subset</b>	<b>All ChEMBL14</b>
<b># of Targets</b>	594	5,272	9,003
<b># of Ligands</b>	2,427	390,866	1,213,242
<b># of Documents</b>	3,394	27,053	46,133
<b># of Points</b>	13,865	948,065	10,129,256

Table 3.3: Number of targets, ligands, and documents in different subsets of ChEMBL 14.

Judging solely by the numbers of targets, ligands, and documents present, the binding affinity subset of ChEMBL 14 appears to subsume a large portion of the whole database, as it contains entries for 32% of all ligands present, and 59% of both the documents and targets (table 3.3). However, the binding affinity subset only contains about 9% of all the points in ChEMBL 14. This would seem to imply that on the whole binding affinities are measured less often than other properties in the medicinal chemistry literature. This impression is supported by the fact that while only 0.09% of all possible pairings of targets and ligands in ChEMBL 14 have some sort of bioactivity associated with them, only 0.05% of all possible pairings of targets and ligands present in the binding affinity subset actually have a  $K_i$  value reported. Binding assays cover a meaningful portion of both the chemical and biological space contained within ChEMBL, but they do so much more sparsely than activities as a whole.

For the paired binding affinity values, it would appear that there is relatively less coverage of all possible binding affinities. Only 11% of the targets with a measured  $K_i$  binding

affinity had at least one biological replicate reported. Similarly, only 0.6% of all ligands present in the binding affinity subset had a biological replicate. In total, 1.4% of the total activity values present in the binding affinity subset were part of a biological replicate binding affinity pair.

The distribution of the ratio of ligands to targets is presented in Figure 3.3. Subplot A reflects the binding affinity subset of ChEMBL 14, while subplot B was created from the paired-affinity difference data. The former curve presents a mean value of 179.8 ligands tested per target and a median value of 26 ligands per target. The maximum number of ligands tested against any single target was 6,789, with 203 targets having had at least 1,000 ligands assayed against them (4.3% of targets), 1424 targets having had at least 100 ligands assayed ( 27%), and 3515 having had at least 10 ligands assayed (66.7%). For subplot B, representing targets with ligands that had been assayed multiple times in the medicinal chemistry literature, 8.27 ligands were replicated on average for each target. The median number of replicated ligands, however, was 3, suggesting, as above a marked skewness in both distributions. This can be verified by visual inspection of the two curves. The overall shape of the two distributions is similar, suggesting that the gross distribution of the paired affinity difference data is not particularly different from that of the binding affinity subset.

The number of times affinity values for a given ligand are reported (alternately, the ratio of papers per compound) are summarized in Figure 3.4. The binding affinity subset data are shown in subplot A. From this distribution, a mean value of 1.17 and a median of 1 are obtained. While at least one compound has been assayed for binding affinity 248 times, 21 have been assayed at least 100 times, 91 have been assayed at least 50 times, and 873 have been assayed 10 times or more, in total this represents less than one-third of 1% of all compounds in the binding data subset. In total, over 90% of the compounds present in the binding data subset have been assayed exactly once. In subplot B, on the other hand, the

paired affinity difference data, by definition, is going to be skewed towards the presence of more reports of affinity for each molecule, and this is, in fact, the case, with a mean number of reports for each ligand included of 3.29. The median value, however, remains at 2, its minimum possible value, and the ligand with the maximal number of independent binding affinity reports had 105, yielding a narrower range of possible values. In this and the following figure, the jagged appearance of the density curve in both B subplots is an artifact of the KDE bandwidth selection algorithm interacting with the low total number of data points and the tight distribution of those points.

The inverse distribution, number of distinct ligands per document, is considered in Figure 3.5. With subplot A again representing the binding data subset, a mean of 18.8 compounds with a corresponding median of 14 is obtained. This is yet another example of a long-tailed distribution, where the papers with the most distinct ligands present had 648 in total. There were 193 papers with more than 100 ligands described, and 1430 that have more than 50 ligands, but these represent only 5.3% of all primary reports from the peer-reviewed literature covered by ChEMBL. Considering the paired affinity-difference data in subplot B, when a paper contains an independent biological replicate value, there are 2.35 repeated compounds on average. The median number of replicated compounds when they occur is only 1, however. The paper with the most replicated ligands had 48 ligands included, and 60 papers had more than 10 replicated compounds, together representing 1.7% of documents present).

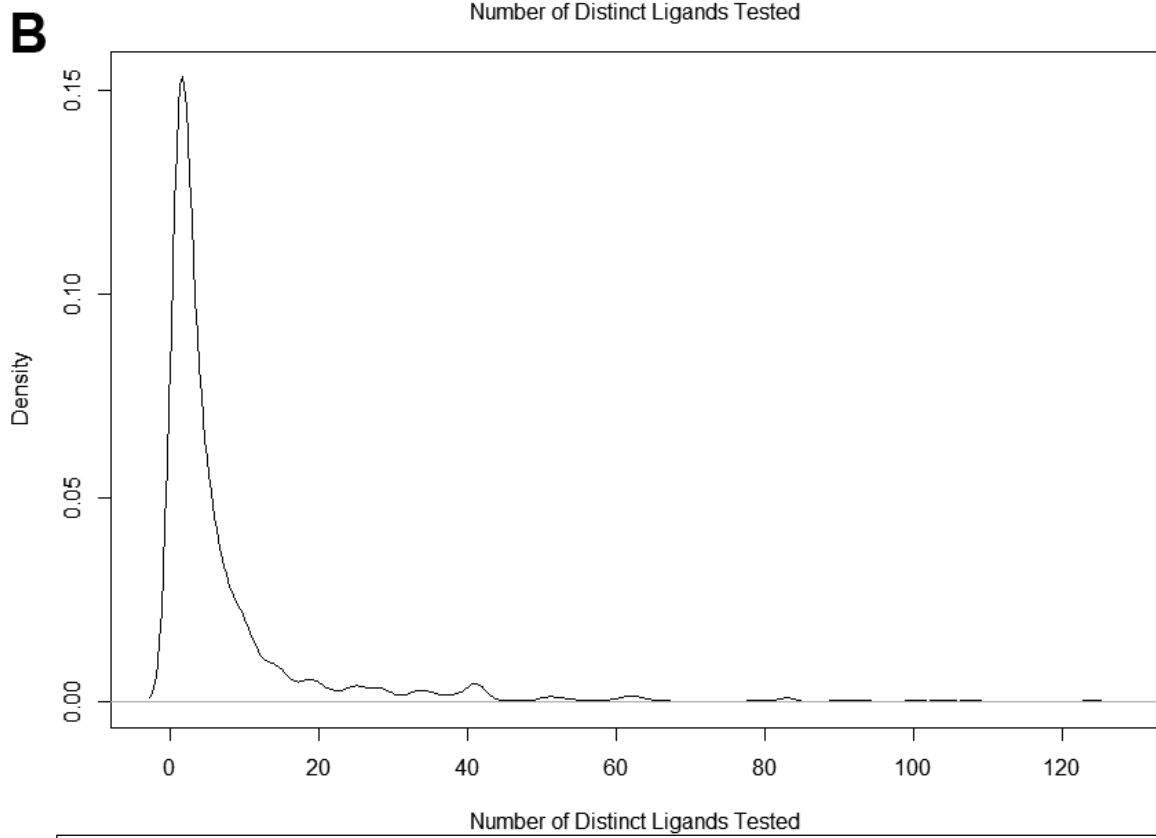
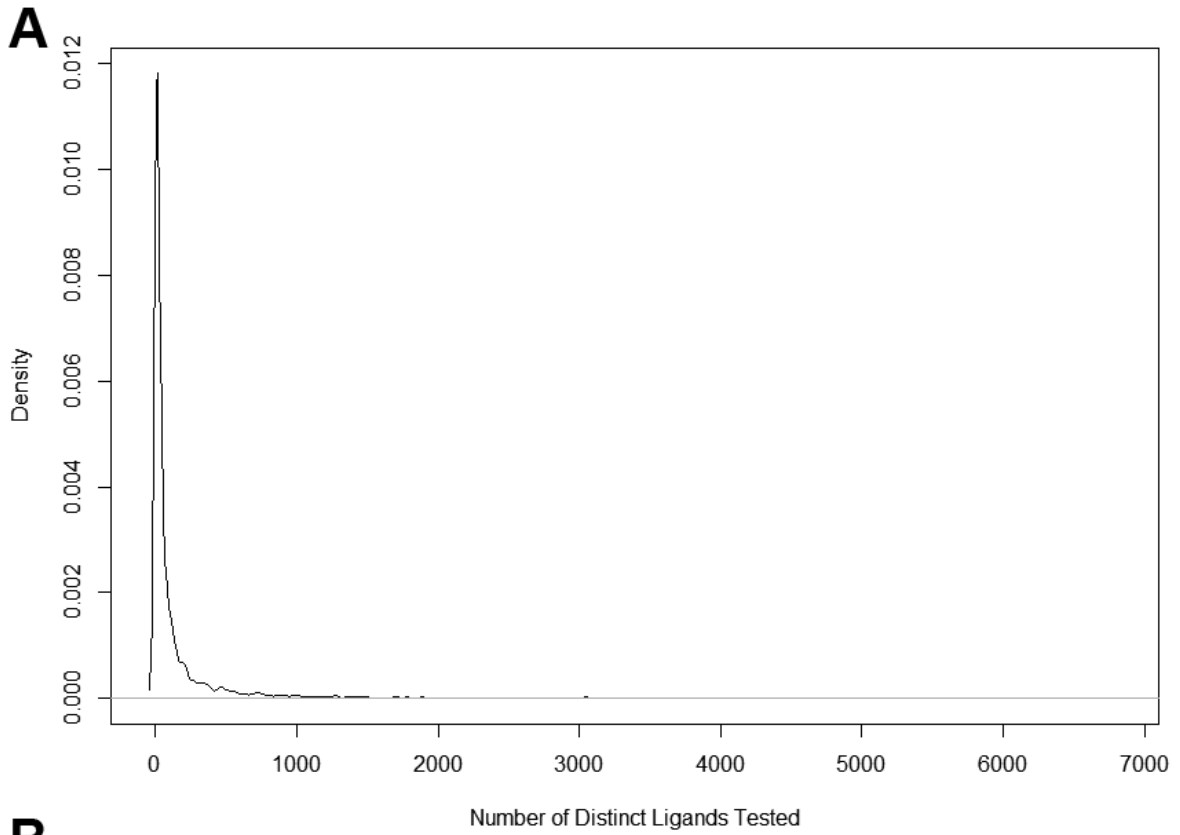


Figure 3.3: Number of ligands tested against individual targets

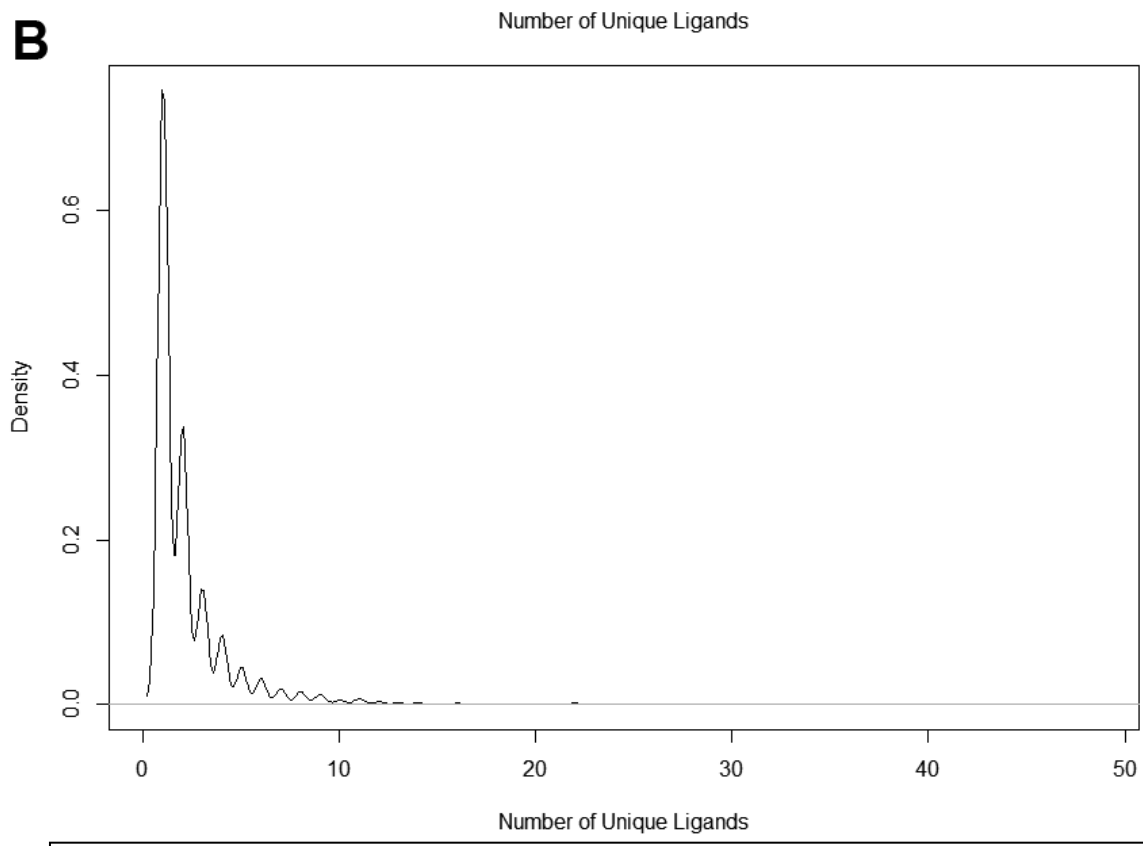
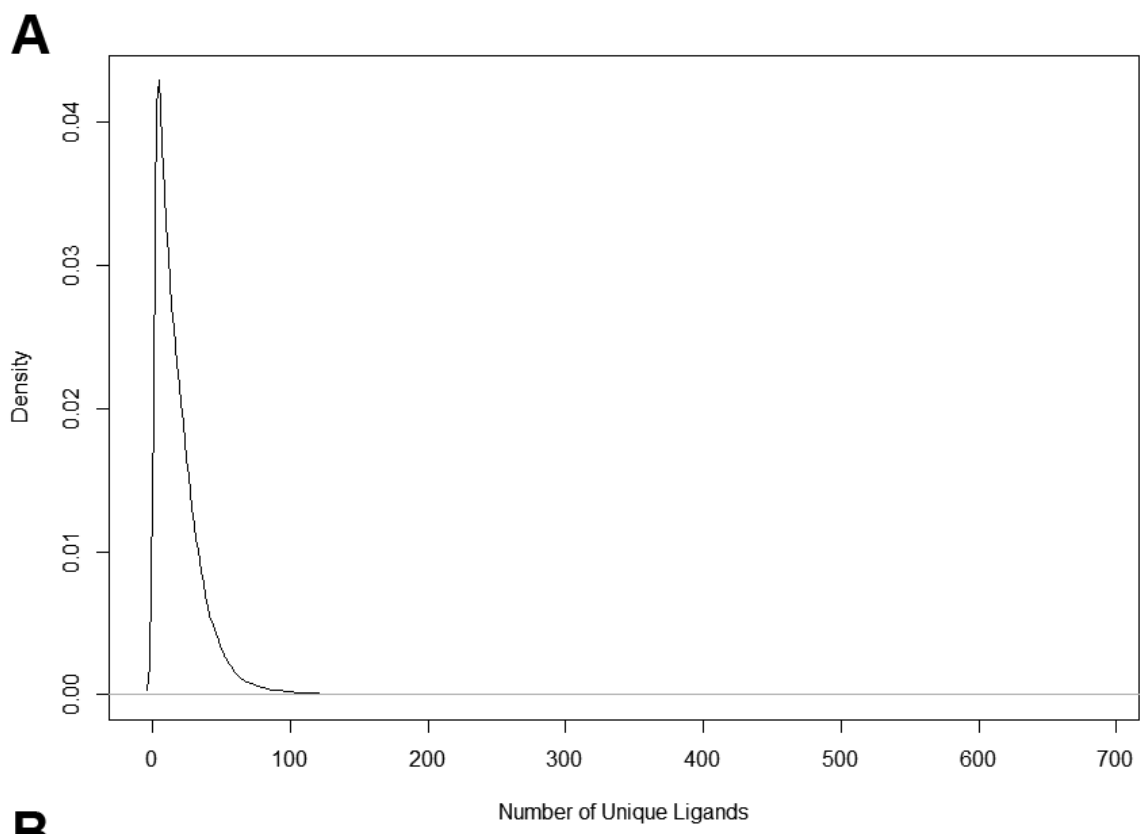


Figure 3.4: Number of times different compounds are referenced in distinct documents

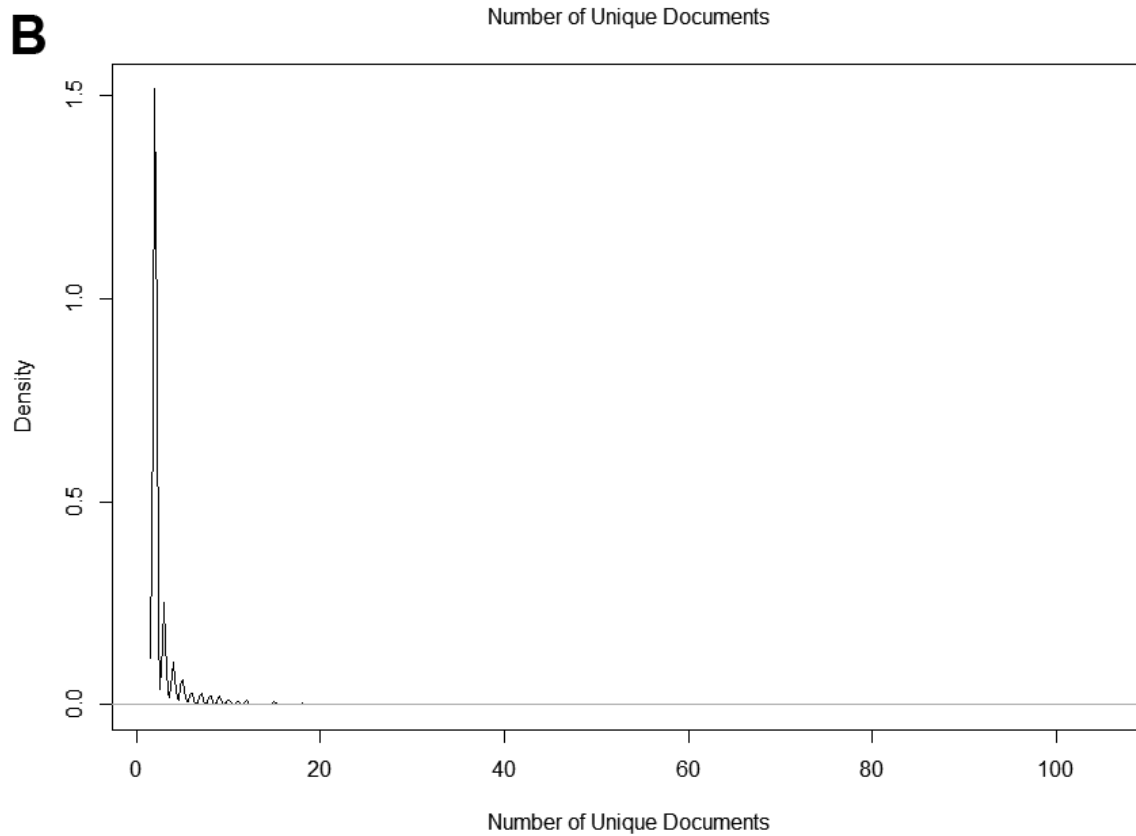
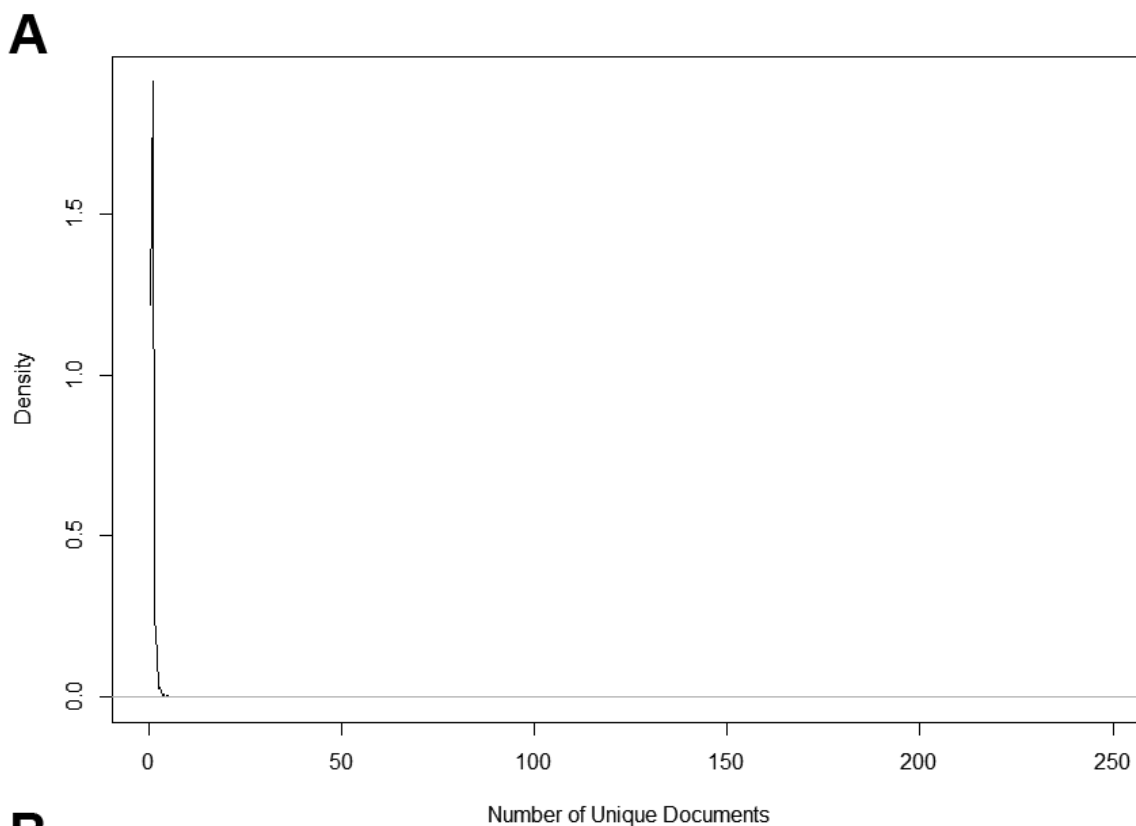


Figure 3.5: Distribution of number of distinct compounds that are present in different documents



*Error types extracted from primary examples*

<b>Number Identified</b>	<b>Error Type</b>
14	Non- $K_i$ values used as $K_i$ units
14	Non-standard Concentrations used
47	Measurement type to units mismatch
57	Modelling study with no original data
9	No units or sources provided
7	Mixed $K_i$ and $pK_i$ values
14	Missing subtype or species information
19	Sign errors arising from $-\text{Log}(K_i)$ vs $\text{Log}(K_i)$
3	Values not found in original paper

Table 3.4: Counts of errors found in a small subset of problematic entries in ChEMBL 14

A non-random sample of 186 papers from the primary literature were taken from documents identified in activity pairs where the difference between the two activity values could not be automatically resolved from the information automatically extracted from ChEMBL (such as negative  $K_i$  values being reported, no units being reported with a  $K_i$  value or units being reported from a  $\text{Log } K_i$  value, or multiple values for a single target-receptor pair being reported) or when the difference between the two  $pK_i$  values were more than 12 log units apart. An analysis of these pairs is summarized in Table 3.4. The predominant source of problematic data in this set is a consequence of data being copied from its original source into papers where computer models are constructed, and then being reintroduced into ChEMBL when those models are abstracted. The second most common discrepancy resulted from failure to import the appropriate units from the original source. Less commonly, but still representing a third of the errors, were issues with  $pK_i$  values

(extraneous negative values, mislabeling  $K_i$  as  $pK_i$ , utilizing non-standard units for determination) and mislabeling or omission of species information for the target. Finally, a nonzero number of errors arose from mixing  $K_i$  and  $pK_i$  values, having no notation as to units of analysis, or the data not being present at all in the cited paper. Thus, altering the process for data acquisition to address only one type of issue, *e.g.*, excluding papers consisting only of modeling data or stricter standards in the original literature (requiring all values to be expressed in Molar units), will not suffice to fully eradicate the identified problems.

#### *Error estimate between biological replicates*

After the removal of all affinity differences of less than 0.05  $pK_i$  units, there are 32,008 data points remaining in the affinity difference data set extracted from ChEMBL 14 (Figure 3.2). With a mean of 0.97  $pK_i$  units and a median of 0.58, it seems improbable that the individual points are normally distributed. In fact, these data are not in a Gaussian distribution, and they are skewed markedly towards smaller values. Given the large number of points available, rather than systematically test different distributions for quality of fit, the use of an empirical cumulative distribution function (ECDF) was an efficient mechanism for describing the distribution of distances between biological replicates which were selected from ChEMBL. A plot of the ECDF appears in figure 3.6. From this plot, it can be seen that over 90% of the differences are less than 3  $pK_i$  units from each other, and that 87.4% are within 2, with 69% separated by 1  $pK_i$  unit or less.

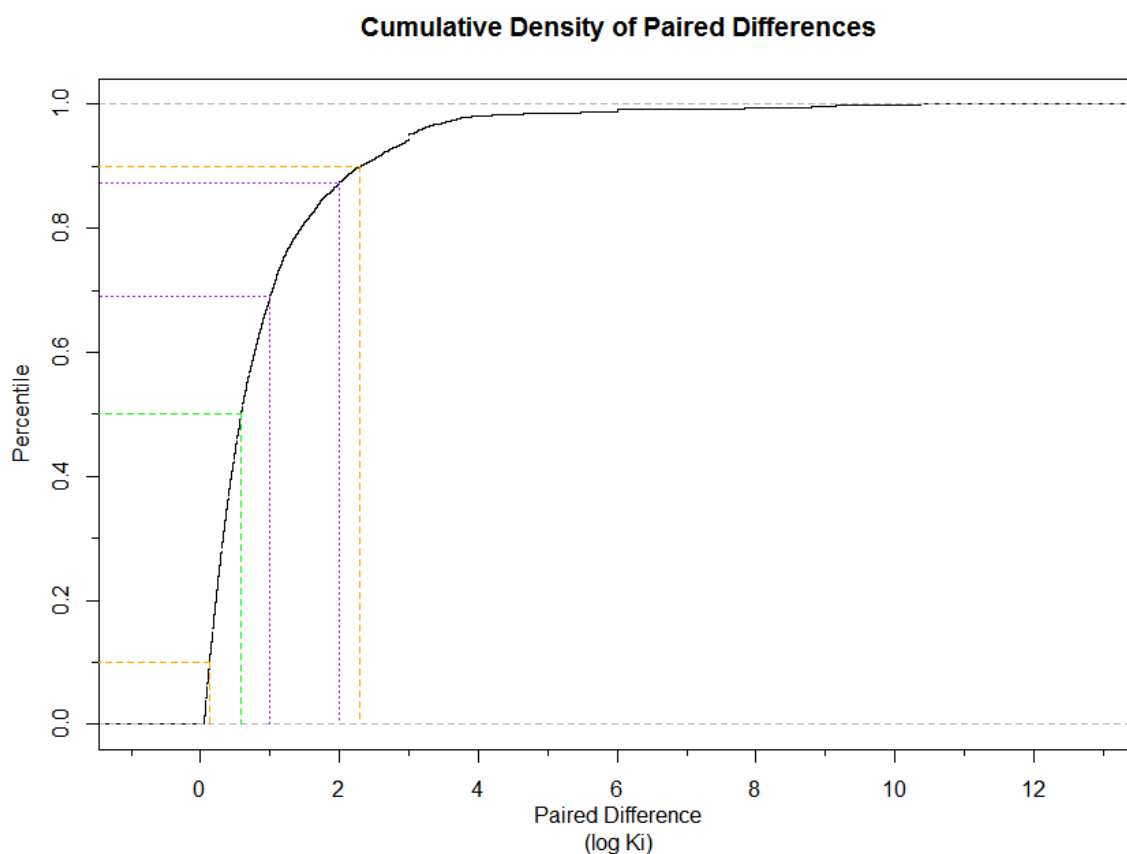


Figure 3.6: Empirical Cumulative Distribution Function of observed differences in biological replicate pairs. Orange dashed lines = tenth and ninetieth percentiles. Green dashed line = median. Purple dotted line = differences of 1 and 2  $\log K_i$  units. (not shown: mean = 0.97  $\log K_i$ )

At this juncture, it is important to note that these calculations have all been based upon the **differences** between two binding affinities and not on the difference between one experimental binding affinity and a standard, error-free, exact value for that binding affinity. As discussed earlier, this hypothetical value does not exist. In order to estimate the uncertainty for a single measurement, we will assert that for each pair of measurements, there exists some value for the binding affinity which, when used as part of the training data for a QSAR model in lieu of either of the paired binding affinity values, will minimize the overall, total error of the model. This value need not be unique for any of the binding affinity pairs; we only assert that one or more values meeting this condition must exist. We

can then use this value as a proxy for the “best” binding affinity value possible for any of the calculated differences in experimental replicate pairs.

If we consider two biologically replicate binding affinities,  $a$  and  $b$ , and the hypothetical optimal error-reducing value for that ligand-target pair,  $\delta$ , we can construct a contingency table that summarizes the relative positions of the three points (Table 3.4). With no *a priori* knowledge of the actual values for these values, we can only examine whether  $\delta$  is inside the interval  $[a,b]$  or not (subject to the definition that  $a \leq b$ ).

	$a \leq \delta$	$a > \delta$
$b \leq \delta$	Outside $[a,b]$	Inside $[a,b]$
$b > \delta$	Inside $[a,b]$	Outside $[a,b]$

Table 3.5: Contingency for position of  $\delta$  relative to  $a$  and  $b$  in hypothetical affinity pair

In half of the unconstrained, possible cases,  $\delta$  will fall between  $a$  and  $b$ . Naturally, this depends on where the experimental points fall on the probability distribution of all possible outcomes (a potential future project would involve investigating the probability of  $\delta$  falling within  $[a,b]$  as a function of the distance between  $a$  and  $b$ ). For the meantime, we will treat this assumption of localized and evenly distributed error as an adequate approximation of the “true” error.

Since these data are being used to compute a mean value (the average error of a binding affinity in ChEMBL) we will always be including both the errors associated with binding affinities  $a$  and  $b$  in the calculation. If we assume that  $\delta$  is inside the interval  $[a,b]$ , we can use the mean of  $a$  and  $b$  to approximate the distance between  $a$  or  $b$ , and the optimal binding affinity value  $\delta$ . This works because any error ( $\epsilon$ ) made in the prediction of  $\delta$  will be canceled out by an equal and offsetting error in the other half of the paired binding

affinity measurement. That is: if the absolute error between  $a$  and  $\delta$  is underestimated by  $\epsilon$  (that is to say that the correct absolute error for binding affinity  $a$  is  $\delta - a - \epsilon$ ) then the correct absolute error for  $b$  would be overestimated by the same amount ( $b - \delta + \epsilon$ ). The average distance between these two points is  $\frac{1}{2} (\delta - a - \epsilon + b - \delta + \epsilon)$ , or  $\frac{1}{2} (-a + b)$ , with both  $\delta$  and  $\epsilon$  cancelling themselves out.

Taking these assumptions as valid and sufficient for a first order approximation, we estimate the mean absolute error of binding affinity measurements in ChEMBL 14 to be 0.48 pK units, with a median of 0.29 pK units. This compares favorably with the values presented in the Novartis paper, which proposed a mean absolute error as 0.44 pK units, with a median of 0.34. Similarly, this suggests that 87.4% of all binding affinities presented in ChEMBL 14 are within 1 pK unit of the “true” value.

## Conclusions

The quality of a QSAR model is implicitly tied to the quality of the data used in the construction of that model. While a large majority of the truly independent biological replicate binding assays appear to lie within 1 pK unit of each other, there are multiple reasons for concern. First, the long tail of large differences in supposedly replicate values, going out to 12 pK units and beyond, suggests that there are enough large magnitude errors in values to cause grief for unwary modelers. With over 90% of all binding affinity values having no independent validation, there is no clear-cut mechanism for identifying these problematic values, even though they must be identified. Ironically, the other major problem appears to be the recycling of data from useful published data sets into computational modeling papers which are then abstracted and reentered into databases such as ChEMBL repeatedly. This tendency to repeatedly recite data leads to multiple hazards. Most significantly, this repetition tends to make the assay values appear to be completely accurate and cast in concrete by repeatedly showing the exact same value. The mere accident that a

modeler found a paper's data useful for a project should not lead anyone to think that it is inherently more trustworthy than any other source. This is not to suggest that ills of open bioactivity data can be laid at the feet of cheminformatics and molecular modeling. Nevertheless, the community has contributed to the problem, particularly in the use of non-standard units for convenience in some studies, and as major consumers of these data sources, it is inherently self-interest to take proactive roles in data quality initiatives.

Beyond the issues already described, several other noteworthy sources of error have already been described. Not all of these are necessarily matters of accuracy in the reporting of bioactivity data. Many of them are meta-data issues relating to the information about an activity value that allows it to be properly classified and contextualized with other similar data. For example, mis-annotation of species, receptor subtype, and/or experimental conditions all contribute to inappropriate bioactivity values being included in the training data for a specific target. Misclassifying data so that inconsistent values, such as affinity ratios or Hill coefficients are comingled with valid affinity data, is another mechanism by which bad biological data is inserted into models.

Those who publish the original papers are not immune from blame, either. While standards for disclosure of experimental details have become much more stringent in recent years, there are still papers which are vague about important details of their experimental procedure, ranging from not discussing the details of expression systems or tissue preparations used to extract target molecules to actually providing binding affinity values (not log values or pK units) without any units at all. The easily visible spikes in affinity differences at exactly 3, 6, and 9 pK units in Figure 3.2 can partially be attributed to the use of nonstandard measures by modelers and errors in data extraction, but the prevalence of these problems suggests that some portion of them are due to researchers being sloppy in their handling and curation of metric units. One final issue that is not necessarily addressed

in current best practices is failures to describe or discuss the provenance of data being used in model building. While most reports in the primary literature provide an immediate reference for third-party data being used, this citation may or may not be the actual initial report of that data; this can lead in extreme cases to citation chains four or five papers long which hamper efforts to validate the original information.

Finally, there are parts of the problem that are surds and which would not be eliminated by better process or more extensive auditing of new information added to data repositories. For example, advances in knowledge that make old data inaccurate or incorrect, such as the discovery of new receptor subtypes, or identification of new binding sites for known ligands in tissue samples. Similarly, no amount of rechecking data will eliminate the structure of modern research which heavily rewards new results over the validation of old results, leading to the situation where there is only one reported binding affinity for over 90% of the distinct small-molecules included in ChEMBL. Lastly, the role of experimental error cannot be underestimated. A poorly conceived protocol can lead to inaccurate results that are undetectable until someone else attempts to replicate it or reproduce results obtained by using it. In view of recent attempts to demonstrate the reproducibility of key findings in high impact factor journals, this last issue is obviously a problem, but it is not one that can be solved by more careful attention to the handling of post-publication data.

In spite of all of these problems, there remains cause for cautious optimism in hoping for improvements in data quality. In more recent editions of ChEMBL, the compilers have already begun two noteworthy changes which address issues that have also been noted herein as sources of error. Activity values for target/ligand pairs which exactly match an already entered data point (value and units) are now being flagged as potential duplicates. Also, activity values of any sort which can be expressed in terms of a concentration are being internally converted into negative log units and provided in data tables along with the

published and standardized concentrations. This does not eliminate error, obviously, but it does suggest that entries that have unphysical values associated with them will be reexamined in light of the failure to calculate the  $-\log$  value automatically.

*Implications for curation of SAR data/Heuristics for curation*

Given this state of affairs, it would not be surprising that oral tradition has sprung up around how best to filter and preprocess data for inclusion in QSAR models<sup>9-12</sup>. Some of this “folk wisdom” has been collected and codified in print as best practices; nevertheless, there are still differences of opinion between practitioners about what constitutes valid practice. In conversations with multiple cheminformatics practitioners, several general themes and suggestions emerged:

- Use no more sources than necessary.
- Favor SAR series papers over HTS data.
- When there are multiple binding affinity measurements, the best value might be near the mean of all the results, but recited data can skew this one way or the other.
- Anything less than  $n$  pK units difference is not worth spending the time to reconcile (where  $n$  is some arbitrary, small value that varies by target, application, and modeler).
- If it seems likely that a classifier model will be the final solution, check for paired values across the classification threshold, even if the difference is less than one pK unit difference.
- Papers reporting a single, novel compound are not particularly useful.
- It is necessary but not sufficient to have consistent data (structural representation, units and significant figures for activities) in order to build valid models.
- It is more effective to verify and curate structural data and activity data at the same time instead of checking structures and activity data sequentially. This is because problems with structural data may be first identified as inconsistency in activity data, and vice versa.
- Similarly, once the data are represented by consistent structures and in consistent units, major errors will frequently appear as a discontinuity in the response variable - that is, highly similar compounds with activity values differing by more than two



orders of magnitude. These points may or may not represent a true activity cliff, however we can expedite the curation process by checking those points first.

In light of these suggestions and the published best practices, we can propose a workflow for structural and biological curation of activity data that can help improve the efficiency and efficacy of pre-modeling data inspection.

### *Proposed Workflow*

If the variability in protein-ligand affinity measurements between completely independent assay providers is such that any difference of less than an order of magnitude, is there any possible benefit in attempting to pick a “correct” affinity value to use for any given molecule? If anything within 1 pK unit of the mean is even remotely equi-probable, it is possible to argue that the choice shouldn’t matter at all, and that regression models in general are probably a waste of time; maybe all that can be accomplished is to categorize molecules by which of a discrete set of activity bins they fall into. While classifier models are still the correct solution, or the only possible solution, for many cheminformatics models, regression models remain useful and relevant, and a curation process for selecting which affinity value out of a set of several reported values to use is still highly relevant. However, the final goal of the selection process may change slightly.

Just as some experimental results are more reliable than others (even if we struggle with identifying which ones those are), not all experimental predictions are equally reliable or useful. While a simple binary classifier can predict whether a given compound is likely to score as mutagenic in an Ames or mouse lymphoma assay that information, however useful, is rarely the end of the inquiry. Even when the classifier is used to screen a large virtual chemical library, having one or several chemical structures still leads to more questions. Most commonly, the questions will take the form of “What can we change to abolish that toxicity?” or “How can the binding affinity to our target be improved?”. While it is certainly possible to optimize compounds by selecting a compound, making a prediction, modifying

the selected compound, making a prediction for the modified compound, and repeating that process to exhaustion or boredom, it isn't always efficient. If all that is available is a binary classifier model, then that may be the best that can be accomplished. With a regression model the same protocol can be followed, but with the advantage that each change in structure has a numerical affinity associated with it. It becomes possible to directly assess the impact of each modification *vis a vis* other compounds, and not just against an arbitrary benchmark. That is, it becomes possible to attempt an optimization of properties, and not just increase the number of structures to be considered.

Such an optimization approach does not even strictly depend on having highly accurate numerical properties. It is certainly possible to attempt an optimization knowing only whether a change will improve or worsen the property under optimization, or whether no significant difference is predicted. But, even when predicted values are known for both compounds, it is the difference in their values that is most immediately useful. An accurate affinity value that closely matches the experimental assay of choice is certainly preferable, but as long as the relative affinities across a set of compounds are well-reproduced, attempting to optimize properties is feasible. Having an estimate of the absolute affinity that is accurate to an order of magnitude will obviously help, but it is the relationship between the different compounds that is primary.

Given that we have estimated the uncertainty of a single compound reported in ChEMBL to be about 1 pK unit, this would seemingly be a death knell for QSAR regression models. This is not the case, however. While most compounds only have their affinity reported once in the public literature, they are not reported in a vacuum. In particular, most reported affinities occur within the context of a paper that reports affinities for more than 10 compounds at once; almost all of these series are performed in a single research group, by the same assay protocol. By definition, the systematic error of these compounds should be

very close to constant and the error present in the difference of the affinities of any two of them will be dominated by the random error in each determination. Current best practices in pharmacology laboratories seek to reduce random error in replicates of affinity measurements to 10%. While this goal is not always achieved in practice, 10% random error in an affinity is significantly easier to work with than an uncertainty of an order of magnitude or more. In light of this analysis, it would seem that a more suitable solution to the identification of the proper activity values to be used in QSAR models would be seek to minimize the total uncertainty in the pairwise differences for the compounds selected by maximizing the number of compounds which were assayed under the same conditions, and preferring values performed in laboratories which have more experience with the assay protocol in use. This is not to suggest that there is not a place for rejecting obviously erroneous activity values, but it seems unlikely that cherry-picking individual data points is sufficient to compensate for the large uncertainty associated with binding affinities singly reported in the literature. Reproducing the differences in affinity accurately allows QSAR models to remain relevant to the lead discovery and optimization process even when data from multiple sources is required.

*Step 1: Standardize chemical structures*

As described previously, the chemical structures in the data set need to be standardized early in the curation process (see Chapter 2). Because we will be relying upon those structures to identify duplicated compounds and values instead of textual annotations, it is imperative that these representations are normalized before other comparisons utilizing those structures are made. A set of standardization rules, either implemented as smirks transformations in OpenBabel<sup>13</sup> or as rules for ChemAxon Standardizer<sup>14</sup> are both acceptable methods for implementing this.

*Step 2: Compare activities for ligands measured at close homologs of the target protein*

In light of the observed problems with species annotation in ChEMBL and the problems in distinguishing assays based on cloned cell line expression systems from native tissue preparation extracts, it is important to verify that reported activities are all measured in truly equivalent systems. Therefore, the activity values for common reference ligands should be checked between the target of interest, and those same ligands as measured against the homologs of the target in other species. Ordinarily, it is sufficient to note whether or not any identical values have been reported for divergent species. Situations where this would be an insufficient condition can be envisioned, however. If a nominally curated dataset does not give rise to predictive models, returning to the primary sources to verify the biological details of the assay conditions is time well spent.

Compound	Target	Species	Assay Type	Value	Units	Date
TLA-042	nAChR $\alpha 7$	Human	$K_i$	2700	nM	Jun 1999
TLA-042	nAChR $\alpha 7$	Mouse	$K_i$	950	nM	Oct 1995
TLA-042	nAChR $\alpha 7$	Rat	$K_i$	1460	nM	May 2002
TLA-042	nAChR $\alpha 7$	Chicken	$K_i$	950	nM	Feb 2000

Table 3.6: Simulated data for affinities of an antagonist binding to nicotinic acetylcholine receptor  $\alpha 7$  in four different species. The exact replication of the mouse and chicken values is a warning sign that the data in the more recent paper may have been improperly compiled or inappropriately copied from the earlier paper.

This kind of error is illustrated in Table 3.6. A tool compound which is useful as an inhibitor of nicotinic acetylcholine receptor  $\alpha 7$  has been assayed for affinity at that receptor as expressed in four different species (human, mouse, rat, and chicken). Everything appears in order, except that the affinity values for mouse and chicken are exactly the same. Because the mouse data was published well before the chicken, it is possible that the authors of the

latter paper used the value published in the former (either out of ignorance, or because it probably was the only published value when they began the research they were reporting). Alternately, the results from the second paper could have been mis-transcribed somewhere in the data assembly process and the species or numerical value is incorrect. Finally, it is improbable, but possible that the affinities for the two different species are identical. In any case, there is no way to resolve this situation solely by reference to the data as presented.

Similarly, there have been situations where mis-annotations have occurred with regard to the specific subtype of receptor being quantified in databases or where mixtures are improperly annotated as being of one specific subtype. In order to guard against this possibility, when data exist for a single compound against multiple variants and subtypes of a receptor in a single species, comparing the value of interest to two or three near homologs and at least one more distant homolog of the same species and class is recommended.

Compound	Target	Species	Assay Type	Value	Units	Date
WDC-331	5-HTR <sub>1A</sub>	Rat	K <sub>i</sub>	420	nM	Mar 1991
WDC-331	5-HTR <sub>1B</sub>	Rat	K <sub>i</sub>	420	nM	Mar 1991
WDC-331	5-HTR <sub>1D</sub>	Rat	K <sub>i</sub>	420	nM	Mar 1991
WDC-331	5-HTR <sub>2C</sub>	Rat	K <sub>i</sub>	11	nM	Nov 2000

Table 3.7: Simulated data for binding affinity for a hypothetical antagonist to various subtypes of serotonergic receptors. The combination of identical affinity value across three variants of a single subtype, species of origin, and publication date strongly indicate that someone seeking to use this affinity data needs to carefully examine the source publication directly.

In Table 3.7, this process is illustrated with a hypothetical serotonergic tool compound that is being considered for inclusion in a QSAR model of the rat <sub>1D</sub> serotonergic

receptor. This entry is troubling for multiple reasons. Primarily, the equality of the three 5-HTR<sub>1</sub> values suggests that there was a transcriptional error either in the initial publication or the data assembly/curation process. Given the publication date, it is also possible that this assay was conducted before the existence of the different serotonergic subtypes were known and the affinity was reported only as 5-HTR<sub>1</sub>. Finally, the species suggests that the sample may have been isolated from extractive tissue preparation and not expressed cleanly in a recombinant cell line. This would have resulted in all three subtypes (as well as several other analogous receptors) being present in the experimental sample with the calculated final affinity being for a mixture of those receptors at an unknown and variable ratio. The only way to clarify which, if any, of these scenarios is the correct one is to examine the primary literature.

*Step 3: Compare values for duplicated ligands at the same target*

Historically, the affinity values for small molecule ligands at their protein targets were determined by competing off a radioactive reference ligand of a known binding affinity. Because this tends to lead to extended networks of interdependent affinity values, consistently using the same values across an entire protocol becomes important for minimizing systematic error. When the same reference ligand is used by multiple research groups for their binding affinity studies, each group will often end up using their own particular value for the reference affinity. Locating instances where the same ligand-target pairs occur at multiple locations in the data set, and then comparing those values is a useful diagnostic. If two different sources report biological replicates with a difference in affinity of less than an order of magnitude, their overall results are likely to be broadly compatible and both includable in a single data set without major difficulties. However, if the difference in replicates is more than an order of magnitude, there may be methodological differences or underlying assumptions in one or both sources which requires the exclusion of one or both

sources from the combined data set. If there are two different biological replicate values reported from the same laboratory, close inspection of the primary report will be mandatory. Either there are subtleties in the reports that were not captured by the compilation process, or a new value for at least one compound in the tested compound was determined in the middle of the work cited. It is important to be certain that the value which is considered most reliable by the researchers reporting the affinities is used. If it is impossible to ascertain which is most trustworthy, it may be necessary to exclude part or all of the data arising from that source.

Compound	Target	Activity Type	Value	Units	Pub. Date
Cordrazine	Procrastin X	K <sub>i</sub>	950	nM	Feb 2004
WDF-209	Procrastin X	K <sub>i</sub>	125	nM	Feb 2004
WDF-314	Procrastin X	K <sub>i</sub>	84	nM	Feb 2004
Dypraxa	Procrastin X	K <sub>i</sub>	2500	nM	Oct 2005
TLA-1138	Procrastin X	K <sub>i</sub>	600	nM	Oct 2005
TLA-1701	Procrastin X	K <sub>i</sub>	51	nM	Oct 2005
Cordrazine	Procrastin X	K <sub>i</sub>	725	nM	Mar 2006
WDF-278	Procrastin X	K <sub>i</sub>	97	nM	Mar 2006
WDF-411	Procrastin X	K <sub>i</sub>	28	nM	Mar 2006

Table 3.8: Simulated affinity data for multiple assays. Two organizations with different reference compounds have published assay results for multiple compounds. The reported value of cordrazine changes from one publication to the next. Before either set of WDF data are used, the reason for this divergence needs to be examined carefully.

In Table 3.8 some of these issues are illustrated. In it, two organizations (Wonder Drug Factory and TLA Pharmaceuticals) have research programs to develop inhibitors of procrastin X and hopefully improve graduate student productivity. The assay used at Wonder Drug utilizes cordrazine as its reference ligand, while TLA's assay is based on

dypraxa, instead. In February of 2004, WDF has a note published which states their determined affinities for cordrazine and two novel ligands of their own devising. 25 months later, when the program has been terminated, another paper is published describing some more selective ligands. However, the affinity for the reference ligand has significantly changed. If this result had come from TLA, it would not have been a major concern. The two values are less than an order of magnitude different from each other. However, the difference between the two values is more than 10%, it does not only represent an increase in the precision of the affinity, and the assays were notionally performed by the same organization which did the previous set. There are plausible explanations for this discrepancy, such as changes in assay protocol or technology, transitioning from a CRO assay provider to an internal laboratory (or vice versa), or possibly even a change in supplier of the cordrazine used. But, without more clarity on this point, it remains unclear whether the two sets of affinities for the novel compounds reported are compatible and should be treated as being from a single organization. A simple correction might be used to make them compatible, or they may need to be handled completely separately, with only one of the two published values for cordrazine included in the model.

*Step 4: Examine fingerprint plots for outliers and activity cliffs*

It is useful to know when compounds in a data set are either extremely different from the majority of compounds present, or when there are compounds that are virtually identical with highly divergent activities. Because a model has not yet been created, the easiest way to assess these properties is to use standard molecular fingerprints and a similarity metric, such as the Tanimoto index<sup>15</sup>, to plot the relative distance between compounds in the data set in a fixed-size space. A more specific approach would be to generate either MACCS<sup>16</sup> or some variant of a connectivity fingerprint for each molecule, take the Tanimoto index over all pairs, and then apply multi-dimension scaling to the matrix of the compliments of the



similarities. The first two elements of the resulting MDS “eigenvectors” for each data point can then be directly plotted on a Cartesian grid and visually inspected. In cases where a single point is several times more distance from its nearest neighbor or where two points are essentially superimposed with a multiple order of magnitude difference in activity, there may be underlying problems with the data.

In the former case, it may be best to remove the chemically distant point, especially if this apparent distance is confirmed by visual inspection of the possible outlier and its nearest neighbors’ chemical structures. This is not a formal rule for identifying outliers such as the Tukey method. The creation of the MDS plot relies on an algorithm that attempts to separate points as much as possible given the similarity/distance constraints placed upon them, and no formal numerical criteria are established for the exclusion of a given data point. Nevertheless, this method has the advantage of being applicable in cases where there is little or no replicate data for a given compound, and it allows for the incorporation of chemical intuition and experience into the workflow. It does have limitations, however. In particular, this procedure should not be applied iteratively more than twice or thrice, and the structures involved should always be visually examined before one or more is removed. By definition, there will always be a compound that is least similar to the majority of compounds in a data set. If the compound or compounds that dominate the spatial distribution of an MDS plot of the computed differences are uncritically and repeatedly removed, it is possible to keep peeling away the outermost points in a data set, layer-by-layer like an onion, until almost all the data are discarded and nothing remains with which to build a model.

In the latter case of two points being close together on an MDS plot but having activities differing by two or more orders of magnitude, a possible discontinuity in the response variable has been identified. If the chemical structures involved are not similar,

there is the possibility that their coincidence in a two-dimensional plot is by chance, and there is a significant distance between the points in higher dimensions that were not captured by the choice of descriptors or fingerprint types. If the two structures are not identical, but are likely to be of a congeneric series or are otherwise structurally similar (based on visual inspection), it may represent a true discontinuity, especially if the compounds were both assayed under the same protocol in the same laboratory. These data points should be flagged as possibly problematic and considered for elimination from the data set if models built with them consistently have problems with low accuracy in that portion of chemical space. If the two activities are not from the same protocol or laboratory, this may indicate either problems with quality control or accuracy in one laboratory or another. This is especially a possibility if the compounds are structurally identical and both assayed by the same provider. In cases such as this, close evaluation of the primary literature is essential.

These issues are illustrated in Figure 3.7. In this simulated plot of the two highest variance dimensions from a multi-dimensional scaling calculation on a data set being assembled for a QSAR study, the structures of 3 compounds designated a, b, and c, are considered. C is a known anti-fungal which is very dissimilar to the other compounds under consideration (which are CNS agents). Based on its relative position in the MDS plot, c should almost certainly be removed. Even if there are other compounds present with pendant carbohydrates or polyene changes, they are still clustered much more closely to a and b than to c, which suggests that they differ from c in some significant manner (probably molecular weight). A and b have the opposite problem. They are practically superimposed in the plot. Examining the structures demonstrates that they differ only by the substitution of

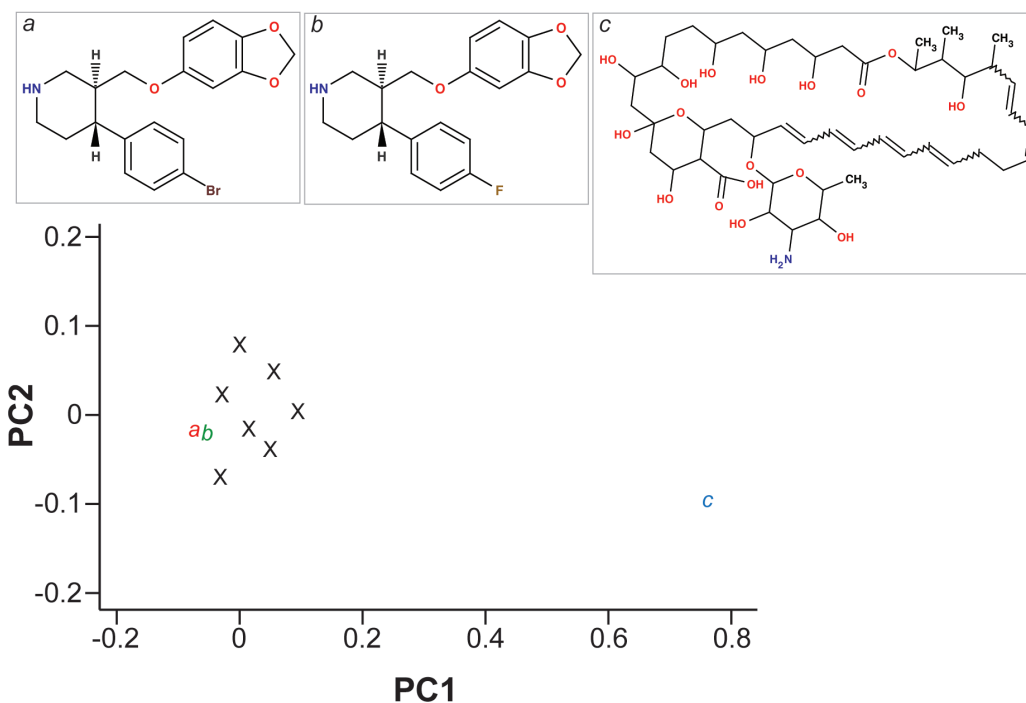


Figure 3.7: Sample MDS plot of several compounds present in a data set. Each character represents one compound. A and B are highly similar. If their activities differ by more than 2 orders of magnitude, they form an activity cliff and one or both may need to be removed from the data set. C is an outlier, both by positioning on the plot and by visual comparison of chemical structures A, B, and C. It probably should be removed

one atom at the same position on a phenyl ring. If the affinities differ by more than 2 pK units, then there is an activity cliff present which needs to be taken into consideration. If the majority of compounds in the vicinity of a and b are similar in affinity to a, then b may need to be removed from the training set; if b is more representative of average affinities in the region, then a should be considered for elimination. Once either compound is removed from the training set, it should not be placed into any external test set as it is already presumed to be a hard-to-predict data point. However, as most QSAR protocols have some form of

variable selection used which may completely transform the similarity relationships between compounds, it may be instructive to apply the final model to the removed compound in order to establish whether that discontinuity in affinity remains, or it was optimized out through the variable selection process.

*Step 5: Spot check primary sources for structural and biological data accuracy*

As has been previously discussed there is a non-negligible level of error introduced into compiled data by the act of compiling it<sup>6</sup>. While it might be desirable to check each primary reference for each compound and activity being considered for inclusion, this is probably not be viable in large data sets with many sources being consolidated. In addition to any primary sources that have been identified as potentially problematic in previous steps, 10 to 25% of the primary sources represented in the consolidated data set being curated should be manually checked at this time, cross-referencing the structures and activities cited in the database to the source tables and figures in the paper and ascertaining that the values being cited are obtained with methods consistent with the other sources, and that there are no special limitations on the results published which would make them unable to be generalized to similar compounds in the same system.

*Step 6: Group and prioritize compounds by assay provider*

As stated previously, the overall goal in this workflow is to maximize the consistency of the data used. While it may be tempting to construct a calibration curve to normalize all affinities to the same baseline, it is rare for more than one or two compounds to be assayed at the same target more than once. This is probably insufficient to construct a calibration curve in a majority of cases. While it may be possible in rare cases, the method currently being proposed can be utilized in cases where there is even less overlap between the different compounds reported in the primary literature, even to the point of there being no compound

that is in common to all assays. But in order to achieve this, all data points being considered will need to be associated with an assay provider. This will usually be the research group that published the original reports of the activity, but caution is needed. An academic research group may have sent compounds to a collaborating laboratory for assay or to a nationally funded resource center such as the PDSP or a MRLN facility for screening. In this case, all compounds screened at the facility should be grouped together, even if they were provided by independent groups. Conversely, if an organization is big enough to have multiple sites that are geographically diverse or has acquired smaller ones, more than one laboratory may have been involved in assaying activities, and multiple protocols might have been used for a single target. In cases such as this, it may be desirable to treat the data as originating from different sources. In this case, the geographical addresses attached to the authors of the primary report in the literature may serve as a clue, as would clear evidence that the compounds reported were being evaluated in divergent therapeutic categories.

If a publication is not a primary report, then it may be best to deprioritize all information contained within, or even to drop it from the data set completely. While a review may prove a source of novel chemotypes that promise to expand the coverage of the model in new directions, the activity measurements will almost certainly be drawn from multiple sources with multiple protocols. If the provenance of these values cannot be ascertained and placed in the context of other results from the same protocols, they may prove to be an impediment to model stability. Similarly, if the activity of a single compound is reported at multiple divergent values (more than 10% variance) by a single assay provider and the problem cannot be traced to errors in the chemical structure depicted, it may indicate that the protocol followed is not reliable or that there are other problems at the provider. Again, it may be desirable to either deprioritize the results or even to eliminate them from the data set entirely. Similarly, it would be appropriate to deprioritize papers which cover molecular modeling or other theoretical and/or retrospective papers at this point. Sometimes novel

values will be included in this sort of paper, either in the form of data abstracted from the primary literature in sources not covered by the consolidated data source currently being used or from new chemical matter being assayed for the first time as a form of prediction validation or extending coverage of an interesting region of poorly defined chemical space. These values should not be excluded *ab initio*, but tracing their experimental sources and assigning them to the correct assay provider is just as critical as with any other data point.

Finally, it is possible that one assay provider or methodology yields results which are at odds with several other sources or that report uncertainties that are much larger than those of other sources. If these errors are sufficiently unsettling, it would be possible to deprioritize or exclude those providers or technologies at this point. This should be done on an all-or-nothing basis, rather than as a reverse cherry pick of single data points that remain untrusted even after completion of earlier steps of this protocol and examination of the primary account of the values. If the protocol or provider is considered a reliable method or source, individual data points should not be cast aside without evidence of error somewhere in the handling of the data.

*Step 7: Select compound value from the highest priority (largest assay provider group) present*

Each compound identified will now have an assay provider associated with each activity value. These providers will ordinarily be ranked or prioritized in the descending order of the number of compounds assayed according to that provider's protocols and instrumentation. Modifications may well have been made for cases where the exact provenance of certain values is unknown or unclear, or when the values from one assaying group are consistently problematic with respect to other sources. For each unique compound present, the activity value that came from the assaying group with the highest priority should be retained in the modeling set; all other values for that compound should be set aside and ignored for this iteration of the modeling process.

### *Step 8: Filter problematic chemotypes*

Once a de-duplicated candidate modeling set has been assembled, there may be compounds included within it which contain chemotypes that are known to be problematic in the QSAR modeling process, either because they contain elements which cannot be included in a given set of descriptors (heavy metals, ionic substances, isotopic or stereochemical isomers, physical mixtures)<sup>17</sup> or are known to be poorly predicted by QSAR methods. These vary from method to method and modeler to modeler, but can include such moieties as quaternary amides, bio-labile esters and amides, and long aliphatic or polyether chains.

### *Step 9: Set aside compounds and values from small assay provider groups*

In cross-validation schemes, the modeling set as a whole is split into multiple folds where each fold in turn is used as an external set to measure the quality of a model built from the other folds. For this to work, the structures in each fold should not be terribly dissimilar from each other. Likewise, it is detrimental to the process for a compound to appear only in the external fold, without anything resembling it in the training set. For this reason, it is desirable, when using an n-fold cross validation scheme, to remove compounds which were taken from an assaying group of n-1 or fewer compounds and set them aside as a second, supplementary hold-back validation set. Strictly speaking, it may be possible to include these results with those from another assay provider using a highly similar protocol, but in practice it can be difficult to determine sufficient detail about both protocols to confirm their congruence.

Similarly, because the value of a compound's inclusion in the modeling set is not just in the absolute value of its activity, but also in its relative and absolute activity compared to other compounds measured according to the same protocol, there is little value of including

compounds that come from an assaying group with no other compounds reported. If the majority of compounds are reported in groups of two or three, a regression model may not be a viable option; using a traditional modeling/external set split, rather than n-fold cross-validation may be more practical.

This workflow will not catch all errors. Nor will it reduce the problem of curation to the point where persons with only a rudimentary scientific background can curate biochemical data. Rather this workflow is intended to catch relatively common and obvious problems, and identify where other issues may exist, so that an experienced researcher can spend their time and effort on stranger, edge case problems and not use their attention span on simple issues that can be identified automatically. There are certainly opportunities for improvement, especially in terms of automation, and improved algorithms for threshold checking of similarity in putative SAR series. Similarly, a more robust analysis of putative activity cliffs would allow for more rapid identification of problematic data.



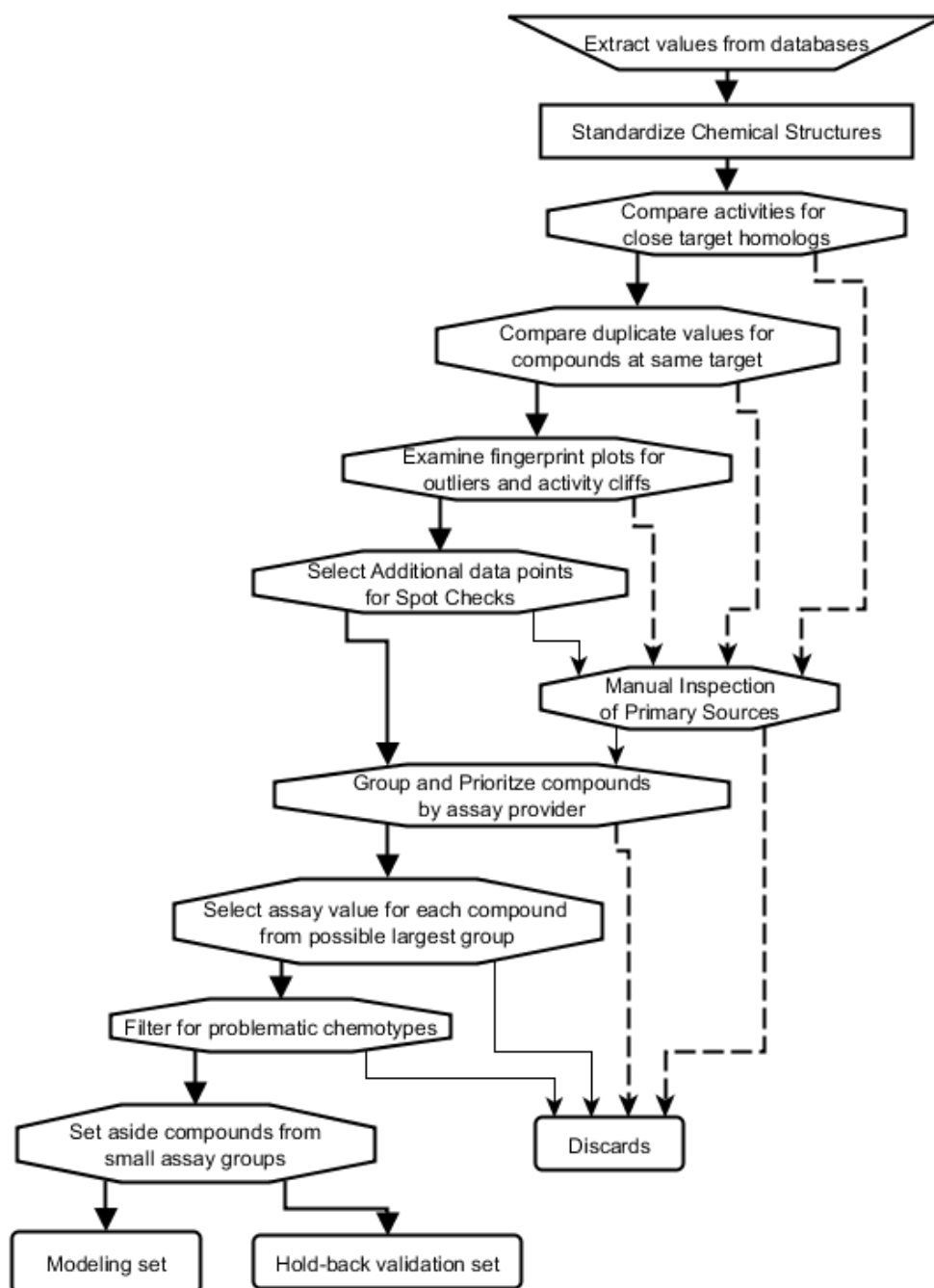


Figure 3.8: Schematic View of Proposed Biological Deduplication Workflow. The primary flow through this system is indicated via the bold arrows from top to bottom. At octagonal steps, a decision is made by the curator. If a data point or compound is consistent with others or non-redundant, flow continues down and to the left. If an inconsistency occurs, flow continues to the right. These data points will require examination of the primary literature to resolve their status, and will then either be dropped from consideration or returned to the workflow at the curator's scientific discretion.

## REFERENCES

- (1) Limbird, L. E. *Cell Surface Receptors: A Short Course on Theory & Methods*; 3rd ed.; Springer: New York, **2005**.
- (2) Maggiora, G. M. On Outliers and Activity Cliffs--Why QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **2006**, *46*, 1535.
- (3) Sisay, M.; Peltason, L.; Bajorath, J. Structural Interpretation of Activity Cliffs Revealed by Systematic Analysis of Structure-Activity Relationships in Analog Series. *Journal of Chemical Information and Modeling* **2009**, *49*, 2179–2189.
- (4) Bento, A.; Gaulton, A.; Hersey, A.; Bellis, L.; Chambers, J.; Davies, M.; Krüger, F.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Research* **2014**, *42*, D1083–D1090.
- (5) Kramer, C.; Kalliokoski, T.; Geddeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K(i) Data. *Journal of Medicinal Chemistry* **2012**, *55*, 5165–5173.
- (6) Tiikkainen, P.; Franke, L. Analysis of Commercial and Public Bioactivity Databases. *Journal of Chemical Information and Modeling* **2012**, *52*, 319–26.
- (7) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science* **2008**, *27*, 1337–1345.
- (8) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, 2014.
- (9) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR & Combinatorial Science* **2007**.
- (10) Dearden, J.; Cronin, M.; Kaiser, K. How Not to Develop a Quantitative Structure-Activity or Structure-Property Relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research* **2009**, *20*, 241–266.
- (11) Scior, T.; Medina-Franco, J. L.; Do, Q.-T. T.; Martínez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Current Medicinal Chemistry* **2009**, *16*, 4297–313.
- (12) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Current Pharmaceutical Design* **2007**, *13*, 3494–504.
- (13) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics* **2011**, *3*.
- (14) ChemAxon, Ltd. *ChemAxon Software*; Budapest, 2014.
- (15) Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **1912**, *11*, 37–50.

- (16) Durant, J.; Leland, B.; Henry, D.; Nourse, J. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Modeling* **2002**, *42*, 1273–1280.
- (17) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50*, 1189–1204.

## Chapter 4: The Accuracy of QSAR Models of $\alpha$ 2a Adrenergic Receptor and Serotonin Reuptake Transporter Binding

*“To do two things at once is to do neither.”*

Publilius Syrus

### **Summary**

Major Depressive Disorder (MDD) is a significant health concern, costing the US economy over \$200 billion a year. The serotonin reuptake transporter (SERT) has proven to be a useful target with relatively high therapeutic index and a low incidence of life-threatening adverse events. While there are several different pharmacological interventions based on the inhibition of SERT that are well-tolerated by a majority of patients, individual responses to these drugs remains highly variable. There are no validated predictors or biomarkers that will identify which drug will yield an abatement of depressive symptoms without intolerable side effects. Protocols exist that enable non-psychiatrist prescribers to initiate first and second choice therapies for MDD with SERT inhibitors and a few related inhibitors of multiple monoamine transporters. However, many patients still require referrals to specialists, either to explore less commonly prescribed SERT inhibitors or to attempt therapy with monoamine oxidase inhibitors or tricyclic agents which both require in-depth management. A second target for MDD therapy has been the noradrenergic signaling pathway, primarily via interference with the norephedrine reuptake transporter. In theory, targeting the individual receptors in the noradrenergic pathway could achieve the goal of mood balance with fewer off-target effects. Based on the biological distribution of the alpha 2a adrenergic receptor ( $\alpha$ 2a) we sought to identify compounds that bind both SERT

and  $\alpha$ 2a as the first step towards identifying SERT antagonist/ $\alpha$ 2a agonist compounds that would ideally show in vivo antidepressant activity.

The identification of compounds with affinity for both SERT and  $\alpha$ 2a begins with the construction of QSAR models for each target. While there are known compounds that show affinity to both targets, a single model for dual affinity is not an attractive solution, owing to both the low abundance of compounds possessing the desired activities, and the origin of those compounds in two previous commercial development campaigns at these targets. Creating separate QSAR models for SERT and  $\alpha$ 2a provides better coverage of the chemical space for affinity at each target, at the risk of requiring more effort to identify ligands with dual affinity. For each target, a raw activity list is extracted from ChEMBL 16 and then curated according to procedures previously described (see Chapter 3). Multiple independent models are created from these training sets, based on zero, one, and two-dimensional topological descriptors using five-fold cross validation, either random forest, kernel support vector machine, or k-nearest neighbors methodologies, each with genetic algorithm variable selection. Models showing acceptable predictivity are retained and combined into a consensus model for each target. Virtual catalogs are structurally curated and pre-screened with molecular fingerprints before descriptors are calculated and normalized for compounds meeting a similarity threshold. Descriptors for the selected compounds are then processed through each component of each model, with a mean score reported for predicted affinity for both targets.

After curation and preprocessing, training sets for  $\alpha$ 2a and SERT had 537 and 2501 compounds respectively. Multidimensional scaling (MDS) scatterplots for each set showed no trivial solutions to separation and no isolated outlier compounds. For  $\alpha$ 2a, 240 models were created with 20 descriptors each. The average correct classification rate (CCR) was 0.81, and no predictive models were identified in y-randomization testing. For SERT, 180

models were created with 25 descriptors each. The CCR for these models was 0.88, with no y-randomized models passing acceptance criteria. Virtual screening of the UNC Center for Integrative Biology and Drug Discovery screening library, Enamine diversity screening library, and Enamine GPCR targeted screening library (approximately 400,000 compounds) were carried out against the models for each target. Eleven compounds were identified as having  $\alpha$ 2a affinity (including one known ligand omitted from the training set), with 429 hits for SERT affinity identified. No compounds were present in both the  $\alpha$ 2a and SERT hit lists, and the compounds identified as potential  $\alpha$ 2a ligands showed limited potential for medicinal chemistry optimization or selective binding affinity. Attempts to construct regression models for  $\alpha$ 2a and SERT were unsuccessful due to low predictivity.

While no usable chemical matter has been identified in this project to date, several other significant outcomes have resulted from this study. Foremost is the development of QSAR models for  $\alpha$ 2a and SERT. Publication of these models will ultimately require experimental validation in the form of binding affinity assays for selected compounds, but that is a matter of patience and new catalogs to screen. A region of low compound density has also been identified in the MDS plot of  $\alpha$ 2a activities, adjacent to a region where several known dual affinity binders exist. This suggests that there may yet be novel chemical entities with dual affinity to be discovered. On the methodological side, the previously described workflow has been shown to lead to predictive models. While more rigorous tests will be needed to quantify the improvement in generated models, it does work. Finally, the entire workflow for the QSAR modeling and virtual screening process has been reimplemented in a high-level scripting language using freely reusable software tools. This is a useful step on its own merits, as it simplifies the modification of the workflow and adoption of new technologies in the Tropsha group as QSAR and machine learning continue to evolve.

## Introduction

Major depressive disorder (MDD) is a major health problem throughout the developed world. It accounts for up to \$200 billion per year in direct, indirect, and social costs in the United States alone<sup>1</sup>. The treatment of MDD was revolutionized in 1986 by the introduction of Prozac<sup>2</sup>, the first clinically successful serotonin reuptake transporter (SERT) inhibitor, or SSRI (selective serotonin reuptake inhibitor), but there is still much room for improvement in MDD therapeutics. In particular, there is often a 4-6 week lag between the initiation of therapy and perceived relief of depressive symptoms by the patient. Also, patient response to different SSRI drugs is idiosyncratic; what is effective for one patient may have no impact on another's mood<sup>3</sup>. These variable responses can lead to extended trials with different drugs as a prescriber attempts to find the correct drug or combination of drugs to provide relief for a patient.

While interfering with serotonin reuptake in the synapse is a validated approach for pharmacological intervention in MDD, the success of non-serotonergic drugs, *e.g.*, bupropion, and multi-target drugs, *e.g.*, nefazodone and mirtazapine, suggests that there are other targets in the central nervous system (CNS) capable of modulating the symptoms of depression. One particular target that has shown some clinical promise is the alpha-2a adrenergic receptor ( $\alpha_2a$ ). In particular, intervention at the  $\alpha_2a$  receptor has shown promise, both to decrease time required<sup>4</sup> after initial dosing for mood elevation to occur and also to limit the common sexual side effects of SSRI therapy<sup>5</sup>. Nonetheless, no compound specifically targeting  $\alpha_2a$  and SERT has been approved for clinical use in the United States, Canada, or the European Union. We set out to identify compounds with affinity for both SERT and  $\alpha_2a$  by the use of virtual screening, using quantitative structure-activity relationship (QSAR) models based on zero-, one-, and two-dimensional molecular descriptors.

There has been a limited amount of work in the peer-reviewed literature regarding QSAR models of  $\alpha_2a$  and SERT binding affinities. In the case of  $\alpha_2a$ , only three papers have been reported which place these two topics in apposition: one paper was not concerned with binding affinity<sup>6</sup>, and a second involved testing a novel method on a data set consisting of binding affinity data for a limited congeneric series of compounds at  $\alpha_2a$ <sup>7</sup>. The final paper was an attempt to construct a CoMFA (Comparative Molecular Field Analysis) model for agonist activity at  $\alpha_2a$ <sup>8</sup>. Unfortunately, this model only presented activities for about 25 previously documented compounds. In addition, CoMFA is a less than optimal choice for our purposes. A CoMFA model is constructed from an aligned ensemble of three-dimensional molecular structures; if the alignment is poor, the creation of a predictive model becomes less likely and any model generated will have decreased accuracy. While it is not difficult to create conformers for a few drug-like molecules of interest and align them to a pre-existing CoMFA model to predict their biological activity, this process does not scale. Thus, CoMFA models are not generally used for virtual screening. As our goal is to find novel chemical matter rather than to expand upon well-defined analog series, CoMFA would not be appropriate here, either.

There are also only a few preexisting SERT models. A targeted literature search returns 12 publications discussing both QSAR and SERT. Of these, two are not primarily concerned with SERT, mentioning it only in passing<sup>9,10</sup>. Another three were actually structure-based models, which are not applicable to our approach<sup>11-13</sup>. CoMFA was the primary technology used in four additional studies<sup>14-17</sup>. Of the remaining studies, one was a methodology study that was demonstrating the ability of decision trees to be used in virtual screening<sup>18</sup>. A second used a small set (47 compounds) of ligands to train an artificial neural network that would discriminate compounds with high SERT affinity from those also possessing affinity for dopaminergic receptors<sup>19</sup>. A final paper, based on classical QSAR



descriptors such as Hammett sigma values and Taft steric parameters, used to examine the SAR relations of two rings in the phenoxyphenyl-methanamine compounds consist of 181 compounds, but the results were reported as IC<sub>50</sub> values, rather than as K<sub>i</sub>s<sup>20</sup>.

Initial attempts at construction of regression and classifier models for  $\alpha$ 2a affinity were not successful. Trials of multiple descriptor sets on Chembench<sup>21</sup> with k-Nearest Neighbors (kNN) and support vector machine(SVM) methodologies failed to reach a minimal level of predictivity (correct classification rate greater than 0.6 or R<sup>2</sup> above 0.35). After this series of failures, the quality of the data was reevaluated.

Over the past decade, it has been demonstrated that errors in chemical structure and in measured biological activity reside in both commercially distributed and publicly available chemical databases<sup>22</sup>. Previous work from the Tropsha lab<sup>23</sup> has demonstrated the importance of accurate data in the construction of QSAR models, suggesting that an error rate of 5% would be sufficient to render any models built non-predictive. This 5% consists of not only errors in reporting activity/affinity and gross errors in chemical structures (ranging from incorrect stereochemistry to misplaced ring substituents, and omitted substructures), but also it includes inconsistency in the depiction of functional groups (such as nitro or nitroso groups) or tautomers, or even duplication of structures with difference activity values. Several papers discussing best practices in QSAR modeling provide guidance for the process of chemical curation<sup>24-27</sup>, but they afford limited guidance into the selection of the correct activity value for a compound at a given receptor from several that may be given in the literature (deduplication). There is also no recommended method for large-scale verification of the accuracy of structures included in a QSAR dataset, only the suggestion of visually comparing computer-rendered structures rendered with the source document(s).

A parallel study into the accuracy of chemical structures from Internet sources relied heavily on consensus comparisons from multiple sources to identify inaccurate structures

and found that manual copying and comparison of chemical structures tends to create errors in chemical structures as much as it eliminates them (see Chapter 2). As over 90% of compounds reported in the medicinal chemistry literature are reported only once (see Chapter 3), consensus structure checking cannot reasonably be used to detect errors in the published structures. With these limitations in mind, it is not feasible to check each structure for accuracy against printed depictions of individual structures. We decided to emphasize consistency of representation and automated structure handling as keys to quality structures and rely on spot checks against printed/PDF reprints to identify problems in the absolute structures.

For any given chemical compound and a set of rules describing preferred depictions, there is one correct representation of that compound. In our case, we do not need to reach that limit of accuracy, only represent the structure in a form that will produce the same descriptor values as the ideal, Platonic structure. For example, when using descriptors that do not consider absolute or relative stereochemistry, such as molecular weight or Kier-Hall indices<sup>28</sup>, the structures used in the calculation may have incorrect stereochemical configurations; this will not affect the accuracy of those descriptors. Likewise, when calculating atom-pair path distances, the nature of the bond between two atoms does not enter into the computation, only the presence or absence of a bond. In this case, issues of resonance structures and aromaticity would not matter for descriptor calculation. This is not to imply that modelers do not need to be concerned with the accuracy of the chemical structures used as the basis of their models, however, there exists a (finite) set of representations of a given chemical structure which will yield equivalent descriptor values. For a given purpose, it may be sufficient and significantly easier to get to a chemical structure that is close enough to a perfect representation than it is to attain a perfect representation.

Beyond structural error, there is also the question of the accuracy of biological activity values. Biological data accuracy represents a significantly more complex problem than structural accuracy. Commonly used biological response variables, *e.g.*, binding affinity or minimum inhibitory concentration, do not have a definitive, correct value that can be measured to infinite precision under all conditions. These values represent a macroscopic ensemble average arising from a large number of individual microstates; their observed values are statistical observations and must be treated as such. In particular, there are uncertainties attached to each value whose magnitudes vary depending on experimental conditions and methods. Conventionally, this uncertainty can be divided into systemic error (the uncertainty in the measurement of the observed value that arises from the inherent conditions of the experimental protocol) and random error (attributed to conditions that are not or cannot be controlled). While it is relatively estimate the magnitude of the random error in a series of measurements (assuming that replicates were run on the measurement and some form of error bar or standard deviation measurement was reported), it is harder to estimate systemic error because it is unusual for activity measurements to be reported for an entire series of compounds from multiple protocols, either from a single research group or from multiple teams of researchers. Typically, only one or two reference compounds will have their activities measured along with a SAR series of novel compounds. This data is enough to verify that values measured are qualitatively correct or even quantitatively reasonable to within an order of magnitude, but it is not sufficient information to construct a meaningful calibration curve that would allow more precise comparison of assay results.

If it were possible to consistently use values from only one protocol as executed by a single laboratory when constructing QSAR models, the problem of systematic error would be greatly reduced. Unfortunately, with current modeling methods, activity values for 50 or more distinct compounds are frequently necessary to construct an acceptably predictive classifier model; a similarly acceptable regression model will often require 100 compound-

value pairs. These values both exceed the mean number of unique compounds reported in a paper in the medicinal chemical literature (see Chapter 3). Even where there are enough distinct compounds described, the nature of drug discovery dictates that most of the data will apply to a small number of series of structurally-similar compounds, with relatively little coverage of most of drug-like chemical space. For common targets, there are larger library HTS datasets such as DrugMatrix<sup>29</sup> and BioWisdom (formerly distributed by BioWisdom, Ltd. Cambridge UK) that can provide sufficient activities for different compounds, but the available values tend to be in relatively well-developed chemical space which may or not be at all similar to the regions available under constraints of intellectual property, target selectivity, and formulation suitability. When building QSAR models from literature activity values, it is highly improbable that a non-trivial model can be constructed from a single source.

In this study, we have applied a novel workflow for the curation of biological activity data to two datasets from the published scientific literature as captured in the ChEMBL database. The final curated datasets contained 527 compounds for  $\alpha$ 2a and 2501 for SERT. These datasets were used to create consensus classifier models utilizing support vector machine, random forest, and k-nearest neighbor methods with variable selection by evolutionary algorithms. The final consensus models both achieved a correct classification rate above 0.8 and suggest regions of chemical space that hold promise for new chemical entities with dual affinity.

## **Methods**

### *Data set extraction and curation*

All structures and activities were extracted from a local installation of ChEMBL (version 16)<sup>30</sup> stored in an Oracle database (version 11g, Oracle Corporation, Redwood Shores, CA). Values reported as being  $K_i$ -based, derived from recombinant human proteins,

and annotated with a sufficiently high ChEMBL confidence score ( $> 3$ ) were retained. In addition, owing to the absence of a collective ChEMBL target ID for mixed samples of  $\alpha 2$  adrenergic receptors, values for  $\alpha 2b$  and  $\alpha 2c$  affinities were also extracted; entries that had identical values for  $\alpha 2a$ ,  $\alpha 2b$ , and  $\alpha 2c$  affinities were discarded as they were typically obtained from membrane preparations of anatomical tissue rather than cell lines expressing pure receptors of a single subtype. Organometallics, inorganic compounds, and counter-ions were removed and variable chemotypes were converted into standard representations according to the protocols described in Fourches, *et al*<sup>23</sup> with Standardizer (ChemAxon, Budapest Version 6.1).

For each affinity measurement extracted from ChEMBL, the source of the measurement was obtained. Compounds and their associated affinities were then grouped according to common assay providers (if any). Each assay-providing group was then ranked in descending order according to the number of unique compounds measured by the group. One review paper reported no original values and cited values for compounds measured elsewhere which were significantly at odds with those reported values<sup>31</sup>. Data from this paper were treated as if the paper had fewer compounds than any other group. Within the results arising from each assay provider, compounds with multiple reported affinities were compared in order to verify the consistency of the reported values. In the event of a discrepancy between two compounds, the more recent (by publication date) affinity value was kept, unless the two values differed by more than one order of magnitude. In that case, both (or all) affinity values for that molecule in the assay group were removed. Assay groups with fewer than five compounds in them were removed from the training set.

Duplicated compounds with independently determined affinity values were identified by ISIDA \_Duplicates (available from <http://infochim.u-strasbg.fr/new/spip.php?article68>). The affinity present in the largest assay group was

retained for use in the training set as long as the difference between it and any other affinity values was not greater than one order of magnitude.

### *QSAR workflow*

After standardization and curation, the resulting data set was split for five-fold external cross-validation. The four splits used for training had their activity values normalized to Z-scores independently of each other (parameters for this normalization were retained in order to normalize external and screening data sets). All descriptors with insufficient independence from other descriptors (defined as having a correlation coefficient of over 0.9) or insufficient variability (fewer than three different values or with a single value occurring at more than 75% of all occurrences) were removed. After generating a pool of candidate models (built using randomly drawn, fixed length subsets of available descriptors and the desired statistical methodology) models were optimized by a genetic algorithm, with the predictivity of each model measured by repeated five-fold cross validation of the training set to determine the mean accuracy or correlation coefficient. At the end of the optimization, the model with the best performance metric that surpassed a selection threshold of 0.7 was used to predict the values of the external fold. If the external selection threshold (0.6) was exceeded, the candidate model was added to the consensus model. After all folds were used in turn as the external fold, the process was repeated to generate additional candidate models with different data splits and/or utilizing different statistical methodologies.

### *Data extraction in R and Caret*

The described workflow was implemented in R<sup>32</sup> (versions 2.15 and 3.02), utilizing the utility functions of the caret library<sup>33</sup> (versions 5.17 and 6.0). By basing the workflow on Caret, the code necessary to generate and manipulate individual folds and to compute model performance can remain constant while multiple modeling methodologies are explored. In

addition, minimal code modifications are necessary to switch between regression and classification models. Finally, there are no proprietary restrictions on R or Caret, ensuring that the source code for derived procedures and the models generated can be shared without encumbering licensure restrictions. The one major disadvantage of this approach is that the implementers of caret philosophically disagree with stochastic variable selection methods and had not incorporated methods for them into their optimization tools. This is not an absolute impediment to using Caret, but it does necessitate either interfacing with another library for stochastic variable selection or self-implementing a method within R.

### *Descriptor generation*

For each molecule in the dataset, two sets of descriptors were generated. Atom-Pair descriptors<sup>34</sup> were calculated for pairs of 10 elements commonly found in bioactive organic compounds (C, N, O, S, P, F, Cl, Br, I, and B) at distances from 1 to 15 intervening bonds. Values were computed using a method described by O'Boyle<sup>35</sup> and implemented in Python<sup>36</sup>, (version 2.7.3), utilizing networkX<sup>37</sup> (version 1.8.1), pybel<sup>38</sup> (version 1.7), and scipy<sup>39</sup> (version 0.13.1) and saved as a comma-separated value (CSV) file. In addition, all available 0, 1 and 2-dimensional descriptors, excepting ionization potential and amino acid counts, were calculated with CDKDescUI<sup>40</sup> (version 1.38) and also saved in a CSV-formatted file. The resulting two files were visually inspected for missing values, merged in Microsoft Excel (version 14.0.7116, Microsoft Corporation, Redmond WA), and re-saved as a CSV-formatted file.

### *Data set splitting*

Multiple folds for external cross validation were selected automatically at random by the appropriate library function in caret. No attempt was made to optimize these splits for diversity or other properties.

### *k-Nearest Neighbors (kNN) modeling*

The kNN QSAR method<sup>41</sup> is predicated upon the assumption that compounds which are geometrically close to each other in chemical spaces are expected to exhibit similar properties or activities. In order to alleviate the problems of sparse data in high dimension spaces, some form of variable selection algorithm is used to reduce the dimensionality of the data set to a more tractable level. Distances were calculated using Euclidian distances (over either all descriptors or a given subset of descripts) and evaluated for optimal predictive power by varying the number of nearest neighbors used to predict the classification or affinity value for a data point from 1 to 6. In classification problems, the majority class of the k nearest neighbors is the class assigned to the data point being predicted; in a regression model, the affinity value is equal to the mean affinity value of the k nearest neighbors. kNN models were implements in R using the functionality included in the caret package (version 5.17).

### *Random Forest (rf) modeling*

Random Forest classifiers were pioneered by Breiman and Cutler<sup>42</sup> and are a generalization of classification or decision trees. A collection of many single decision trees (a forest) is generated by randomly drawing samples with replacement from a training set and then randomly selecting a subset of available descriptors to use as a classifier at each node of each tree. The final prediction for any sample is taken as the consensus value of all of those models. Random Forest models were created using default parameters with the randomForest package of R (version 4.6-7) as encapsulated in the caret package (version 5.17) using default parameters.



### *Support Vector Machine (SVM) classifiers*

SVM methods were first described by Vapnik in the mid-1990s. They generate a hyperplane that separates a set of points in Euclidian space (or alternately in a higher dimension kernel-transformed space) into two classes. The hyperplane is selected by minimizing the number of incorrectly-classified points, while simultaneously maximizing the distance from the point nearest to the hyperplane from each class to the hyperplane. In this work, all SVM calculations were performed using the R package kernlab (version 0.9-19) as interfaced to by the caret package (version 5.17). A radial basis function kernel was used in all models constructed.

### *Model assessment*

Classifier models were optimized in terms of the balanced accuracy or correct classification rate (CCR). Given a data set with two classes a and b, the CCR is defined as

$$\text{CCR} = 0.5 \left( \frac{N_a^{\text{corr}}}{N_a^{\text{total}}} + \frac{N_b^{\text{corr}}}{N_b^{\text{total}}} \right)$$

where  $N^{\text{corr}}$  and  $N^{\text{total}}$  are the number of correctly predicted data points for each class and the total number of data points present in each class in the training data, respectively.

Employing the CCR as an optimization metric for QSAR models allows the use of imbalanced data sets for model creation without introducing a prediction bias into the final model.

Regression models were optimized in terms of the Pearson correlation coefficients of the experimental binding affinity and the predicted binding affinity. Internal predictivity was assessed through repeated five-fold cross validation over the training set. These coefficients, also termed  $R^2$  and  $q^2$ , were calculated in R (version 2.15 and 3.02) according to the methods described in Golbraikh and Tropsha<sup>43</sup>.

## Genetic Algorithm variable selection

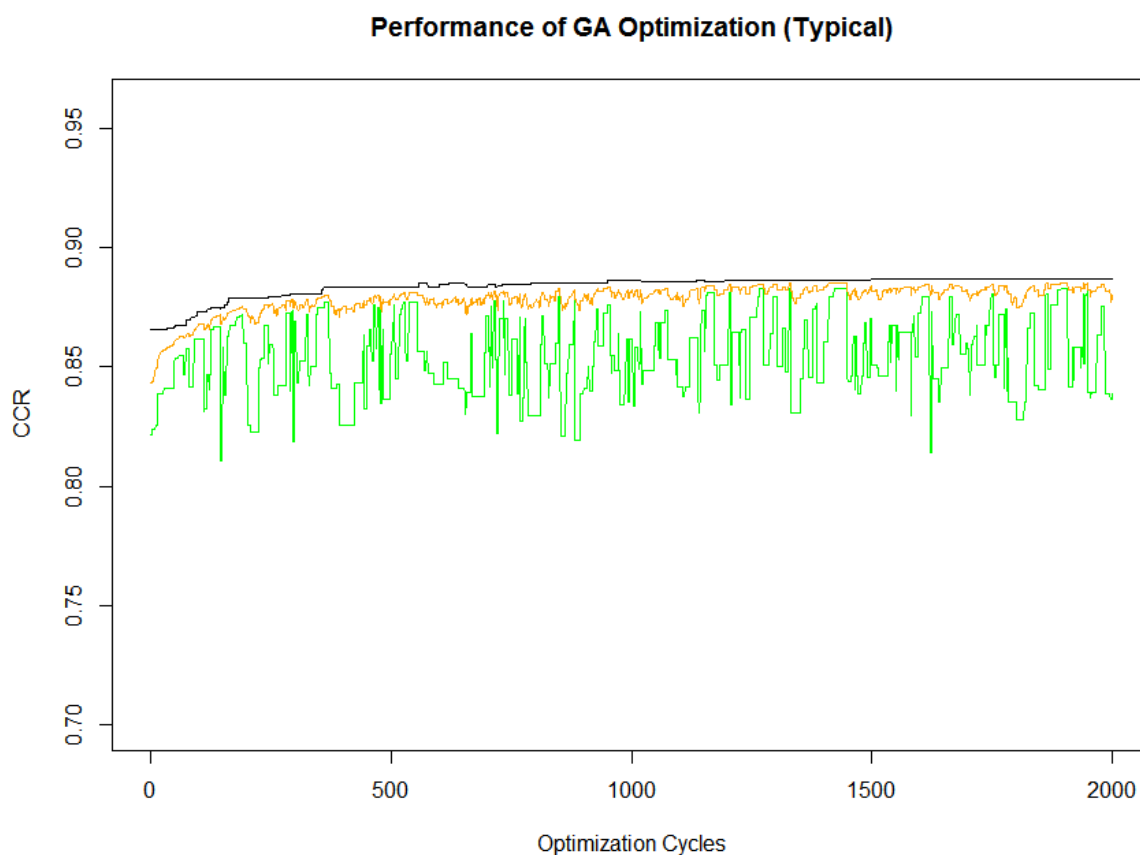


Figure 4.1: Typical plot showing improved performance of GA-selected descriptor subset over time. Black line represents CCR of the best performing model, and orange line represents mean performance of all models. Green line is CCR of the worst performing model.

For this study, a customized genetic algorithm designed to optimize the predictivity of a fixed-length subset of all descriptors was developed. A pool of four to twelve fixed-length chromosomes containing a subset of available descriptors was built and scored using balanced accuracy or  $R^2$  calculated by repeated five-fold cross validation. The initial subset of descriptors was chosen at random. For each optimization cycle, two chromosomes were selected at random. After possibly being mutated at a single position (probability of 1/11), the chromosomes were cut and crossed over at a random position. Any duplicate descriptors

resulting from the crossover procedure were resolved by replacing one of the duplicate bits with a new random bit (potentially repeatedly) until all bits the child chromosomes were unique. The resulting child chromosomes were then scored by determining the CCR value of a model built using only the descriptors associated with the selected bits. The four parent and child chromosomes were ranked by the resulting CCR values, and the top two scoring chromosomes were returned to the model pool for further optimization. After 100 to 3000 cycles (dependent on the type of model being built, data quality, and the size of the chromosomes and the pool) the chromosome with the best performance in the pool of candidates (assuming a score above 0.7) was used to predict the activities of the external fold. Results of a typical optimization run are shown in Figure 4.1. While the quality of the worst model in a pool varies both upwards and downwards as random mutations and crossovers occur, the mean CCR and that of the best performing model both show a consistent upward trend. Peak performance is obtained in this case somewhere around 1000 cycles and no significant improvement is seen after this point.

### *Prediction of activity*

To predict the classification of an unknown compound, all descriptors used in the construction of the models were generated. For each model in the consensus model, the appropriate subset of descriptors were then selected and normalized, before being passed to the classifier. This yields a vector of  $n$  predictions, one for each sub-model. For a negative/non-binder/inactive class, the prediction is coded as zero; for positive/binder/active class predictions, it is coded as one. The sum of this vector is then calculated and divided by the number of sub-models used. This yields a value between zero and one that represents the overall prediction of the consensus model. If the value is less than or equal to 0.25, the unknown is classified as negative; if it is greater than or equal to

0.75, then it is considered positive. Values between 0.25 and 0.75 were considered inconclusively predicted.

### *Virtual screening*

Virtual screening of chemical libraries against consensus models began with standardization of the library structures in Standardizer using the same configuration as used for standardization of the training set. The standardized structures were then pre-screened against the structures of the training set in lieu of a formal applicability domain. MACCS and hashed extended connectivity fingerprints were generated for all compounds in Knime<sup>44</sup> (version 2.8.2) using the CDK fingerprints module, and compared using the Indigo Fingerprint Comparison module. All screening molecules that had a Tanimoto similarity<sup>45</sup> of greater than 0.8 (using MACCS fingerprints) or 0.65 (using extended connectivity fingerprints) to at least one molecule in the training set were collected and had descriptors generated as described above. The resulting descriptors were then subjected to the above described prediction process.

## **Results**

	<b><math>\alpha</math>2a</b>	<b>SERT</b>
<b>Initial Size</b>	1476	3265
<b>After biological curation</b>	669	3194
<b>After rational deduplication</b>	539	2531
<b>After excluding problematic chemotypes</b>	537	2501

Table 4.1: Sizes of data sets for  $\alpha$ 2a and SERT at multiple stages in the curation process

## *$\alpha$ 2a*

Initial attempts at modeling binding affinity were based on data extracted from the 2009 public release of  $K_i$  data from the Psychoactive Drug Screening Program<sup>45</sup> and supplemented with data found in ChEMBL release 2. These compounds were uploaded to the Chembench web server and split into a training set with 70% of the compounds as a training set, 15% in the internal test set and 15% in the external validation set. Attempts to generate predictive classifier and regression models using k-Nearest Neighbors and SVM methodologies were made, but no models were built that met standard thresholds for internal or external predictivity.

### *Extraction and curation of data*

After the workflow in Chapter 3 was devised, data from ChEMBL 16 were extracted and processed as previously described, with data from 38 individual projects incorporated into the final data set of 537 compounds (Table 4.1). A threshold of 6.6 p $K_i$  was set for designating compounds as binding or non-binding. This allowed a ratio of less than 2:1 between the classes. Of particular interest in this dataset were the DrugMatrix HTS binding assay results for the  $\alpha$ 2a receptor and the binding affinities for a second project<sup>4</sup> that had already attempted to optimize compounds for dual binding selectivity against both alpha-2a and SERT. While that project was unsuccessful, the compounds are of particular interest here as they can be used to define an accessible region of chemical space where affinity to both targets can be found.

A multi-dimensional scaling (MDS) plot of the compounds in the final data set is found in Figure 4.2. The distances that form the basis of the plot were extracted using the inverse of Tanimoto indices of 2048-bit hashed extended connectivity fingerprints between all pairs of compounds calculating using the fingerprint library and the cmdscale function in the base R distribution. On this plot, compounds binding to the  $\alpha$ 2a receptor are marked

with circles and compounds that did not bind are marked as crosses; compounds which had high binding affinity for both  $\alpha 2a$  and SERT are denoted with the filled triangles. Green points are those that were taken from the DrugMatrix HTS data, and orange points originated in the Johnson & Johnson Toledo project. Of particular interest, the HTS data points cluster around the upper left quarter of the plot, while the dually-active compounds are mostly gathered around a line near  $y = 0.1$ . This would imply that there are more rigorous requirements for compounds that would bind to both SERT and  $\alpha 2a$ , and that most of the compounds that show up in typical diversity screens are not likely to meet the prerequisites for dual activity. Also, the compounds in the dataset tend to cluster into three separate groups with a moderate amount of separation between each other. There is some overlap between the well-distinguished J&J Toledo compounds and the majority of the HTS screening hits. This suggests that there are additional regions of chemical space within this gap that might be further explored for new chemical matter.

#### *Model optimization for $\alpha 2a$*

In order to assess the effect of the number of descriptors on model performance, subsets of the ensemble of descriptors were repeatedly (30 times for each model length) selected at random and used as the basis of SVM classifier models. The performance of each model was assessed by internal five-fold cross validation without any optimization of descriptors. The results of this experiment are presented in Figure 4.3. In light of the inability of more descriptors to significantly increase the CCR after twenty were selected, the model size for the production models was fixed at twenty.

#### *Production models*

Models were constructed according to the previously described procedure with 20 descriptor subsets of the data set. Variable selection was through genetic algorithm for 3000 cycles with fixed model length. All models constructed exceeded the thresholds for internal

and external predictivity with CCRs above 0.6 and 0.7, respectively. The performance of these models is summarized in Table 4.2.

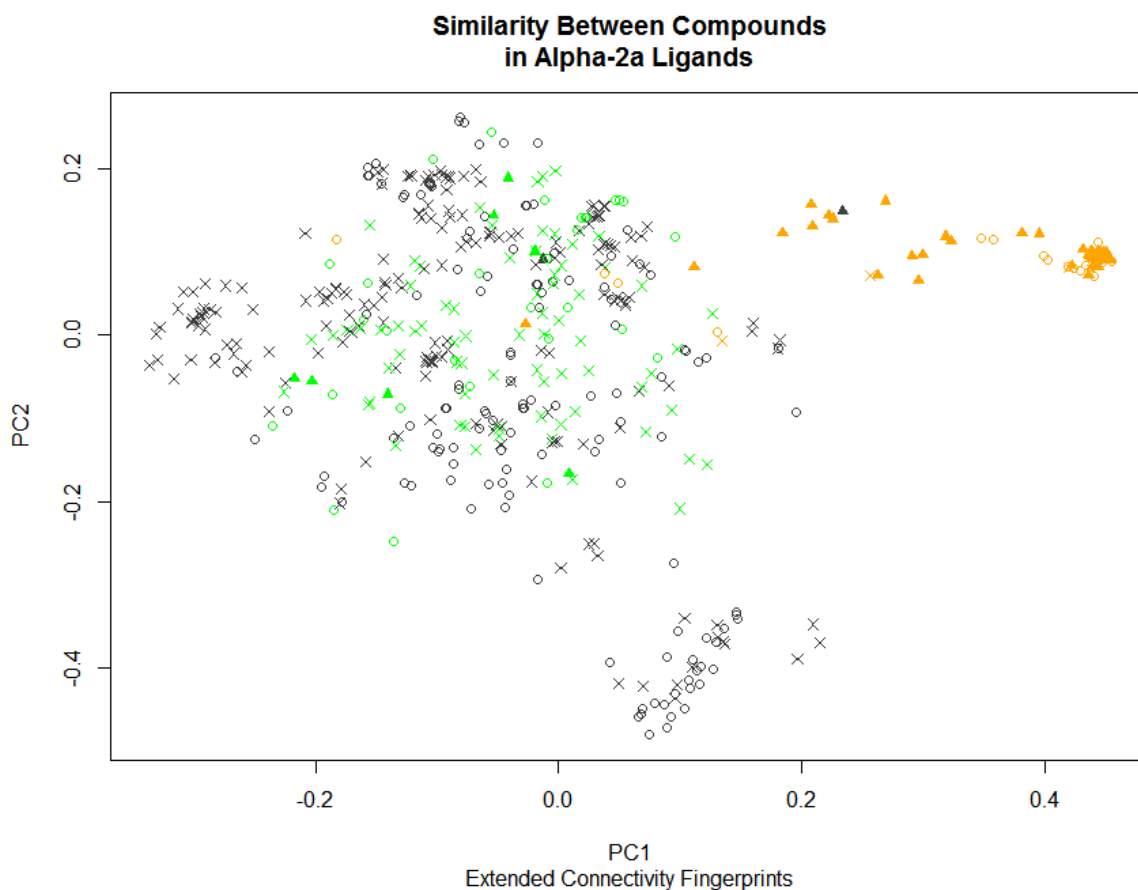


Figure 4.2: MDS plot showing relative similarity of  $\alpha$ 2a ligand set projected into a plane. An open circle (o) represents a compound that binds to  $\alpha$ 2a, An x (X) represents a non-binding compound. A triangle (▲) represents compounds that bind to both  $\alpha$ 2a and SERT. Green points are from DrugMatrix and related HTS screening projects. Orange points are from J&J Toledo work on dual-action compounds.

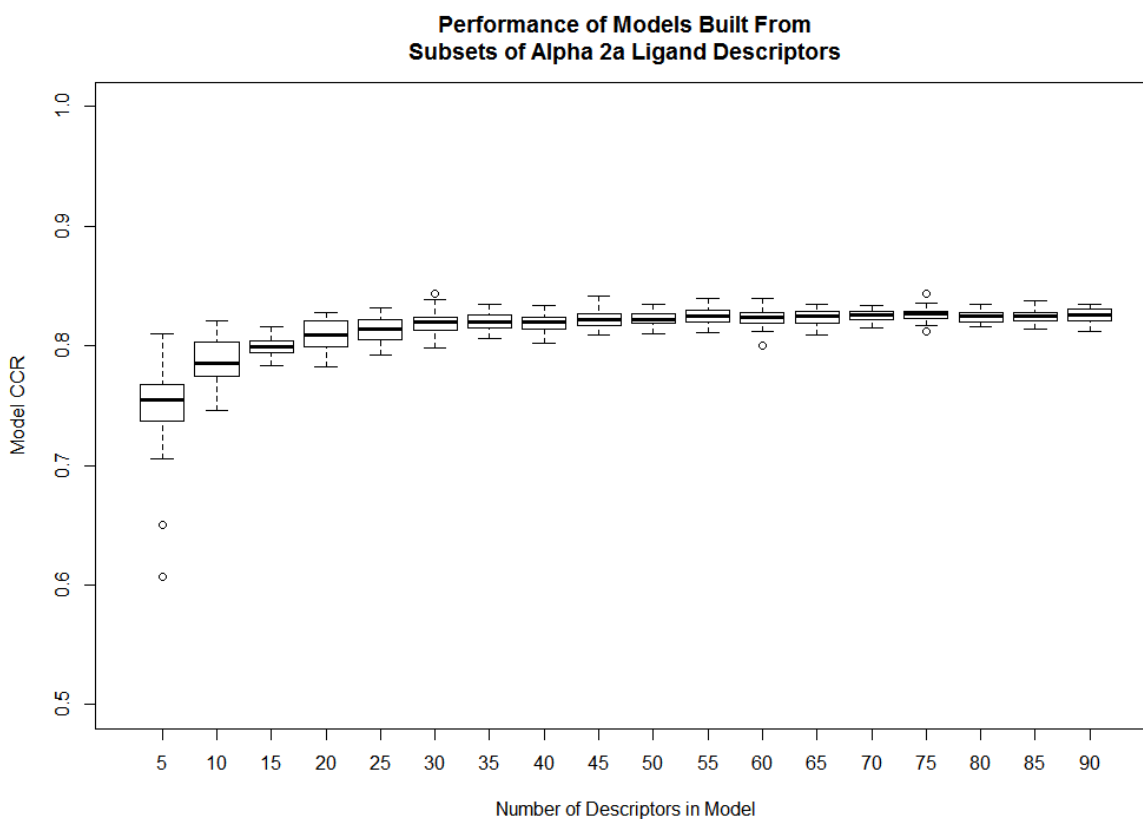


Figure 4.3: Performance of models of  $\alpha$ 2a binding based on randomly drawn subsets of descriptors of various sizes without descriptor optimization.

Method	kNN	SVM radial	Random Forest	Consensus
Number of Models	80	80	80	240
CCR	0.80	0.80	0.82	0.81
Sensitivity	0.81	0.78	0.79	0.79
Specificity	0.79	0.83	0.86	0.83
Precision	0.81	0.78	0.81	0.84

Table 4.2: Summary of  $\alpha$ 2a adrenergic models



### *Y randomization*

Thirty models were constructed from 20-descriptor subsets with *y*-scrambled activity values. After descriptor selection, none of these models had an internal cross validation CCR equal to or greater than 0.6.

### *Regression modeling*

In order to assess the viability of consensus regression models for predicting  $\alpha$ 2a binding affinity, a series of models were constructed using the same, fixed model size GA method for variable selection as was described for the classifier modeling. kNN regression models were created using 6 through 16 and 20 descriptors and 5-fold internal and external cross validation. While some individual models exceeded a threshold for internal predictivity of a correlation coefficient of 0.7, none of them also exceeded the external threshold of 0.6 at the same time and, when averaging over all models of the same descriptor size, no consensus models exceeded either of those thresholds.

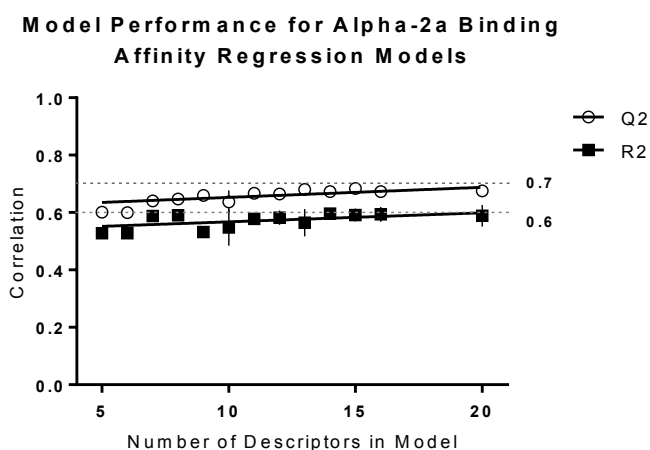


Figure 4.4: Performance of different length subsets of descriptors in regression kNN models of  $\alpha$ 2a binding affinity. Models approached significance in both cases, but never met the designated correlation thresholds (internal data set correlation of 0.7, external validation set correlation of 0.6).

### *Semi-curated data set*

In order to assess the effect of the data curation process, kNN regression models were built from a data set of 702 reported binding affinities for  $\alpha 2a$ . This represents a superset of the smaller dataset used to construct other  $\alpha 2a$  models with no deduplication applied, and some biological errors (affinities for a mixture of  $\alpha 2$  receptors in non-human whole tissue homogenate instead of recombinant receptors in cell lines) retained.

Five models were built using descriptor subsets of length 15, and another five were built using a descriptor subset length of 20. Both models were evaluated by 5-fold internal and external cross validation. The 15-descriptor models missed the 0.7 threshold for internal cross validation with a mean  $q^2$  of 0.693, although they easily passed 0.6 threshold for external correlation with a mean  $r^2$  of over 0.68. One of the 20-descriptor models passed the 0.7 threshold, although the mean  $q^2$  of all 5 was 0.691. The  $r^2$  of the internally predictive model was unfortunately 0.589, although the mean of all 5 models was 0.629.

### *Virtual screening for $\alpha 2a$*

Three different virtual libraries were used to screen for novel chemical matter binding to  $\alpha 2a$ : the UNC CIBDD diversity screening library<sup>47</sup>, the Enamine diversity screening library, and the Enamine GPCR targeted screening library<sup>48</sup>. Each was standardized in ChemAxon Standardizer and then pre-filtered against the training set for the  $\alpha 2a$  models. Compounds that had either an 0.8 Tanimoto similarity (based on MACCS fingerprints) or a 0.65 Tanimoto similarity (based on extended connectivity fingerprints) to a least one compound in the training set had a full descriptor set calculated as described above. These descriptors were then normalized to match the scaling of the descriptors used to generate each model and presented to the saved classifiers. Within the three libraries, no predictions exceeding the 0.75 prediction threshold were made from the first two libraries. In the Enamine GPCR targeted library, 11 compounds were identified that were predicted as

active above a threshold of 0.7 (Figure 4.5). These included nephazoline (E), which is a known  $\alpha_2a$  agonist and which was not included in the training set.

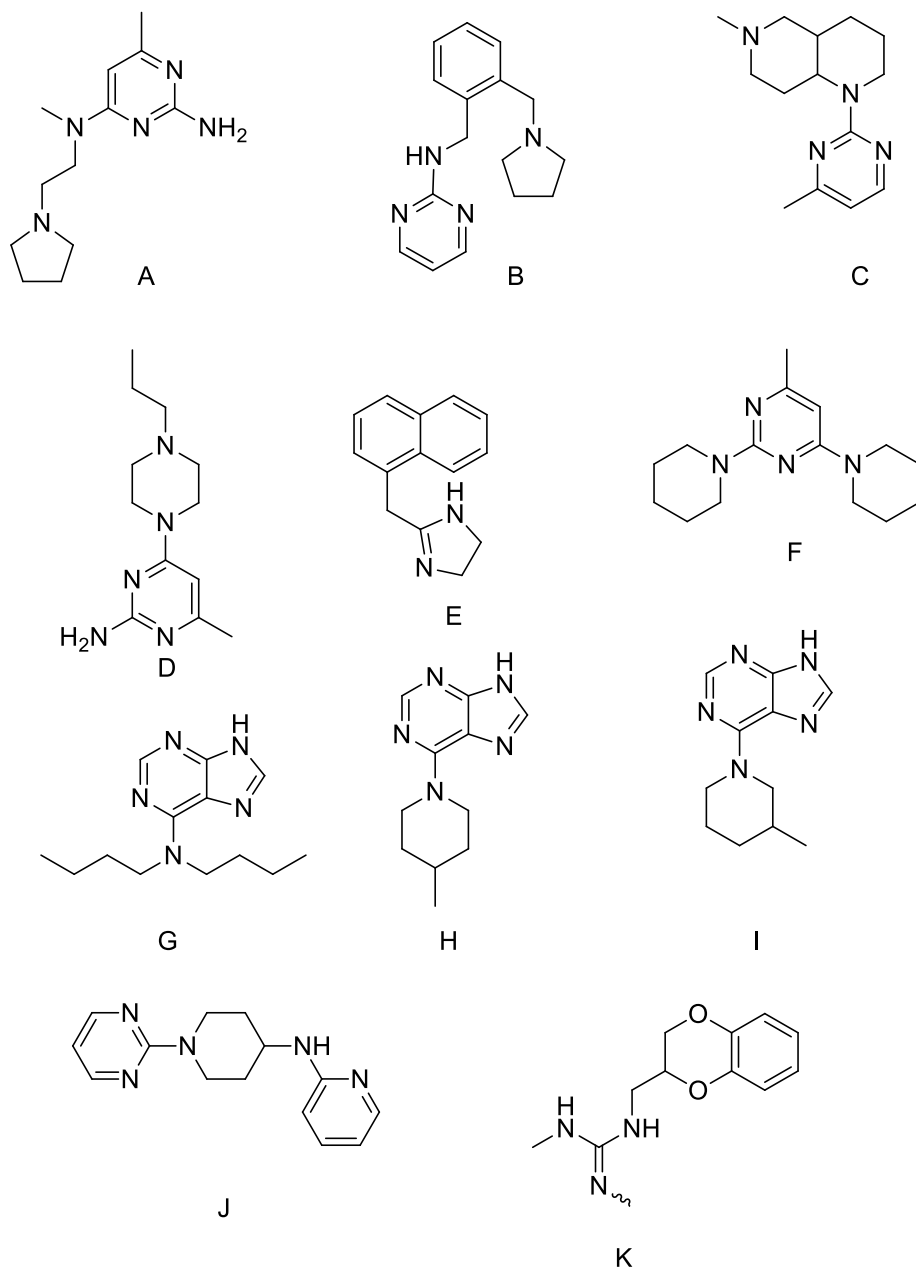


Figure 4.5: Compounds showing possible  $\alpha_2a$  binding affinity from Enamine GPCR targeted screening library. Note that compound (E) is a known  $\alpha_2a$  agonist.

## SERT

### Data extraction and curation

$K_i$  values for compounds showing affinity for SERT were extracted from ChEMBL 16 from 59 different assay providers and subjected to the workflow as detailed above. The initial 3265 data points extracted reduced to 2501 after processing and curation as described in Table 4.1. For binary classification, compounds with a  $pK_i$  of greater than 6.5 were considered binders, and all others were considered non-binders.

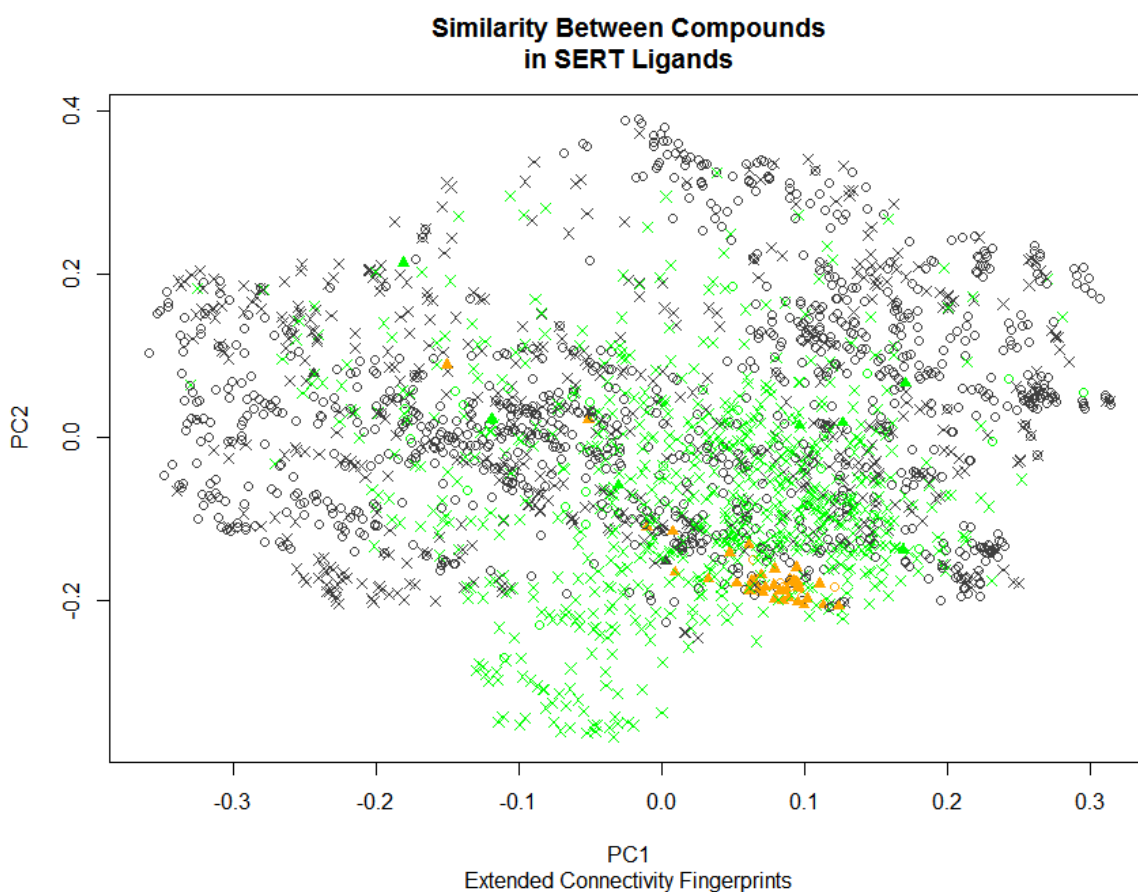


Figure 4.6: MDS plot showing relative similarity of SERT ligand set projected into a plane. An open circle (o) represents a compound that binds to SERT, while an x (X) represents a non-binding compound. A triangle (▲) represents compounds that bind to both  $\alpha_2a$  and SERT. Green points are from DrugMatrix and related HTS screening projects. Orange points are from J&J Toledo work on dual-action compounds.

Construction and examination of an MDS plot (Figure 4.6) showed no obvious sources of trivial solutions for classification and a broader distribution of points than was seen in  $\alpha 2a$ . Data obtained from HTS campaigns were generally distributed similarly to that arising from more focused projects, but there are relatively fewer HTS values at the edges of the graph, except on the bottom edge, where a large cluster of HTS results is placed with almost no data from other sources. In that cluster, there are only a very few compounds that show high binding affinity to SERT.

#### *Classifier modeling*

Similarly to the work performed on  $\alpha 2a$ , a jackknifing process was used to identify an optimal number of descriptors necessary to capture the underlying structure of the SERT data set. Descriptor sets of varying size were randomly drawn from a pool of deorthogonalized descriptors calculated from the data set and used to construct SVM models without any further variable selection or optimization. Thirty models were built at each descriptor subset size and evaluated by their CCR via five-fold internal cross-validation. The resulting models are summarized in Figure 4.7. Based on visual inspection of the upper limit of predictivity at different model lengths, it was determined to use 25 descriptor models to construct a consensus QSAR model for virtual screening.

Models were constructed according to the above procedure with optimized 25 descriptor subsets of the data set. All models constructed exceeded the thresholds for internal and external predictivity with CCRs above 0.6 and 0.7, respectively. The final performance of these models is summarized in Table 4.3.

#### *Y randomization*

Thirty models were constructed from 20-descriptor subsets with *y*-scrambled activity values. After descriptor selection, none of these models had an internal cross validation CCR equal to or greater than 0.6.

Method	kNN	SVM radial	Random Forest	Consensus
Number of Models	60	60	60	180
CCR	0.87	0.88	0.89	0.88
Sensitivity	0.90	0.90	0.90	0.90
Specificity	0.84	0.86	0.89	0.86
Precision	0.84	0.85	0.88	0.86

Table 4.3: Summary of serotonin reuptake transporter (SERT) models

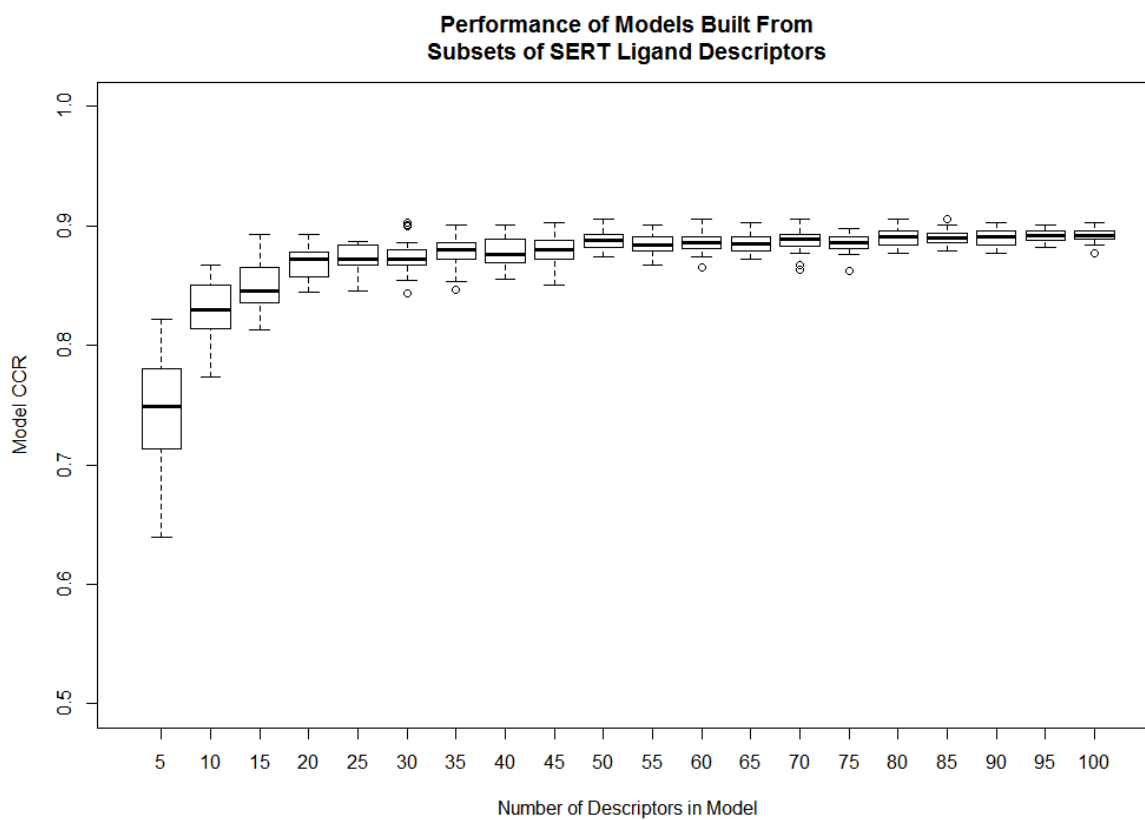


Figure 4.7: Performance of models of SERT binding based on randomly drawn subsets of descriptors of various sizes without descriptor optimization.

Virtual Screening for SERT

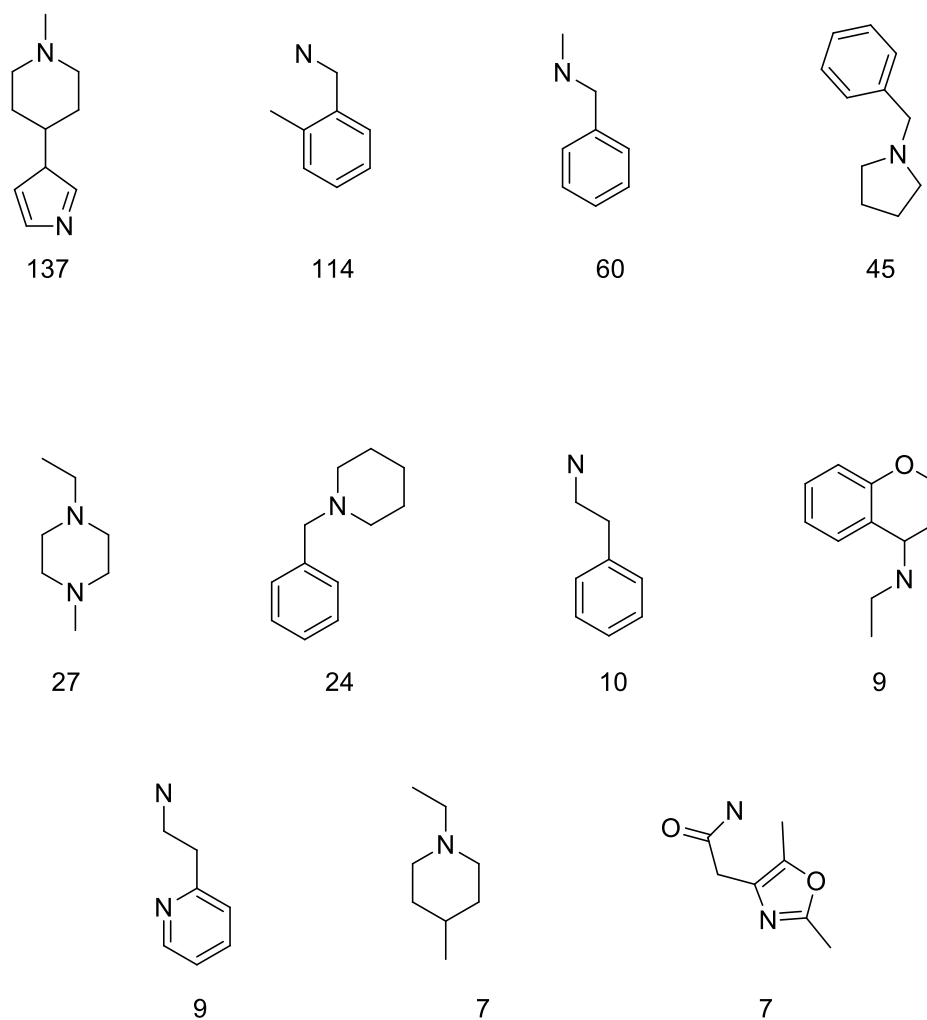


Figure 4.8: Commonly occurring fragments from 429 compounds identified as having potential SERT binding affinity from Enamine GPCR targeted screening library. Numbers underneath the fragments are number of times occurring in hits. No compounds were in both the  $\alpha$ 2a and SERT virtual screening hits.

The contents of the Enamine GPCR-targeted screening library were pre-screened and screened against the SERT consensus model according to the above-described procedures. 429 compounds were identified as potentially having a binding affinity for SERT. None of these compounds were also predicted to have affinity for the  $\alpha$ 2a receptor. These compounds were clustered by maximum common substructures (MCS) occurring by

JKlustor (ChemAxon, Budapest, version 6.01). The most frequent MCS fragments are reproduced in Figure 4.8 with their respective occurrence counts.

#### *Overlap between $\alpha$ 2a adrenergic hits and SERT hits*

As previously stated, there were no compounds that were found to be highly probable virtual screening hits for both the  $\alpha$ 2a adrenergic receptor and SERT. An all-to-all pairwise comparison was undertaken on both sets using modified ECFP fingerprints and the Tanimoto coefficient. One compound from the  $\alpha$ 2a virtual screening hits (Enamine Z128309138, compound B in Figure 4.5) was found to have a similarity of 0.76 to a compound predicted to have high affinity to SERT (Enamine Z219287196, compound A in Figure 4.9 below). Visual inspection confirms that these two compounds are chemically similar. Unfortunately, the conjugated nitrile group in the Z219287196 is known to be a chemical feature associated with aggregation and non-specific assay interference behavior<sup>51</sup>, making it a less-than ideal compound for experimental validation. A second compound from the SERT virtual screening hits (Enamine Z142403862, compound B in Figure 4.9) is less similar to Z128309138, with a similarity of only 0.61. It is however, similar to a marketed drug that targets  $\alpha$ 2a, naphazoline (Figure 4.5, compound E). This is not virgin chemical space. It does suggest, however, that these scaffolds might be useful to identify several compounds that show enhanced selectivity for these two targets.



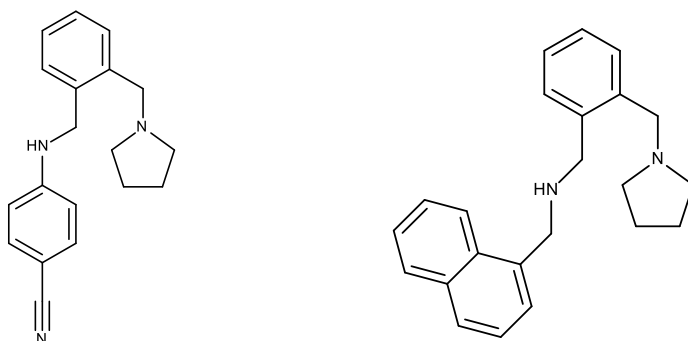
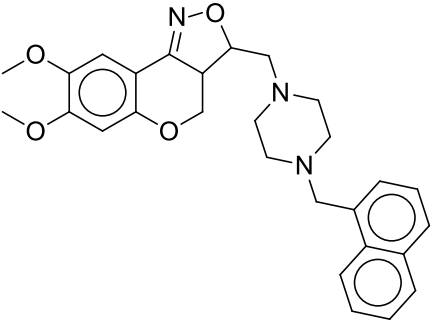
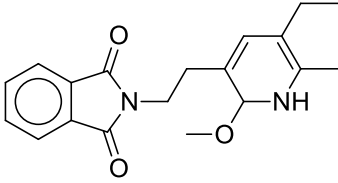
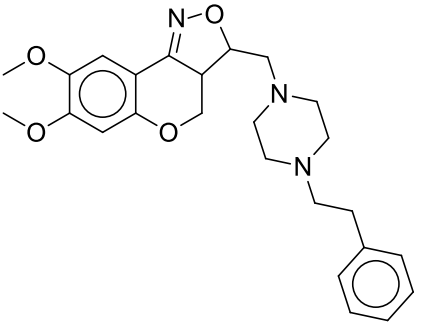
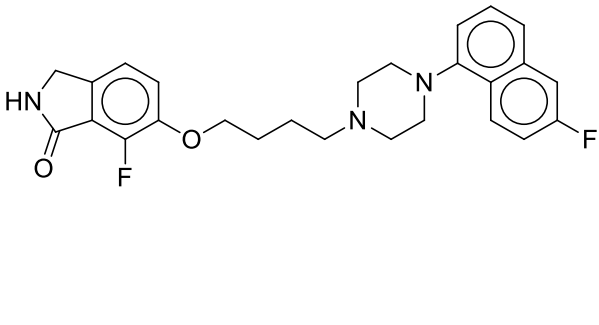
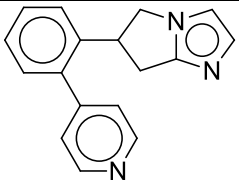
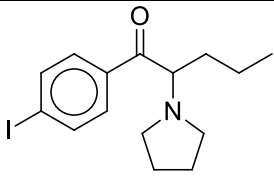


Figure 4.9: Compounds Z219287196 (A) and Z142403862 (B) from the Enamine compound collection are two compounds predicted to have high affinity to SERT that are chemically similar to a compound predicted to have high affinity to  $\alpha$ 2a. These scaffolds are a potential starting point for directed virtual screening or synthesis for dual-affinity ligands.

#### *Consensus model performance on training data*

In order to gain insight into the predictivity of the final consensus models, each training set was predicted by the final consensus model for the appropriate target. This was expressly not done to select any subset of models as being more or less predictive, but simply as an error check in order to be confident that the consensus models could produce meaningful output, and that the above procedure for virtual screening was not so rigorous as to exclude any compound from being identified as a potential screening hit. Based on the 0.25 and 0.75 thresholds used to identify non-binding and binding compounds, only three compounds in each training set were incorrectly classified as being predicted to be in the opposite class from their actual activity (structures are provided in Table 4.4). In addition, 50 compounds out of the  $\alpha$ 2a training set were not definitively predicted as active or inactive (9.3% of the total). For SERT, 186 compounds (8.1 % of the total) were not definitively predicted. In light of the inappropriateness of optimizing against a test set, which in effect the entire training set is under external cross validation, it would be inappropriate to

attempt to use these values as a metric for further optimization of the consensus models, but it does offer some degree of confidence that the consensus models are not selecting compounds as active or inactive at random, and that they can identify some subset of molecules as likely to possess high binding affinity.

<b><math>\alpha</math>2a</b>	<b>SERT</b>
	
	
	
<p>Table 4.4: Training compounds consistently mis-predicted in final consensus models for <math>\alpha</math>2a and SERT. In addition, a subset of training compounds from each set (50 in <math>\alpha</math>2a and 186 in SERT) was not reliably predicted to be active or inactive by the final consensus models.</p>	

## Discussion

Two datasets have been constructed to describe the chemical space that defines small organic compounds capable of binding to  $\alpha$ 2a or SERT. The relative size of the two datasets (the SERT dataset being approximately four times the size of the  $\alpha$ 2a set) implies that there is more detailed information about SERT affinity. This is in line with the relative commercial interest in these two targets. While both models demonstrate tendencies for binding or non-binding compounds to predominate in different regions of MDS plots, a trivial separator between these classes cannot be drawn in two dimensions for either data set. This tends to suggest that neither data set is trivially separable, and therefore of limited value as the basis for a QSAR model. The presence of larger regions in the  $\alpha$ 2a dataset with few or no data points suggests that there are regions of chemical space that have not yet been explored (or at least reported on) for  $\alpha$ 2a adrenergic activity and that may be novel chemical matter available for exploitation.

While the presence of gaps that may be exploitable in the  $\alpha$ 2a space is good news, the descriptors that are most often selected for the constituent members of the consensus models do not offer significant guidance in potential structures to consider. Of the eighteen descriptors most frequently selected to be included in submodels within each consensus model (selected from approximately 145 independent descriptors out of over 2000 total available), only four were present in both models (see Figure 4.5). Furthermore, three of the four descriptors identified are simple carbon types indices which measure the substitution patterns of saturated and unsaturated carbons at primary, secondary, and tertiary carbon atoms. While they are useful for virtual screening in combination with other types of descriptors, they do not provide much guidance for which fragments are likely to be useful in proposing novel structures for a compound binding with high affinity to a given target.

Similarly, the fourth descriptor (third order charge autocorrelation) is a measure of how charge density repeats over groupings of three atoms at a time and is likewise useful in virtual screening, but relatively opaque for rational design purposes. On a more optimistic note, two descriptors that were selected which have intrinsic chemical meaning: N<sub>2</sub>N and N<sub>8</sub>CL. N<sub>2</sub>N is a count of how many times a nitrogen atom is bound to another atom that is bound to a second nitrogen atom (N-X-N in general form). This motif is present in the pyrimidine ring and the guanadyl group noted in all compounds predicted to bind well to the  $\alpha_2a$  receptor. The N<sub>8</sub>CL fragment is not as specific, encapsulating all compounds where a nitrogen atom and a chlorine atom are separated by 8 bonds (or seven atoms). It does not readily suggest any particular functional groups, but it does suggest that a chlorine atom is likely to be associated with SERT binding affinity, particularly when combined with a nitrogen atom located at a distance exceeding a 5 or 6 membered ring. Halogens are not uncommon in the structures of serotonergic drugs; this motif is seen in the structures of sertraline, nefazodone and trazodone, and zimelidine and paroxetine both show a similar N<sub>8</sub>X motif where X is bromine for zimelidine and fluorine for paroxetine.

#### *Deduplication strategy and relevance*

The ability to build models from uncurated data with predictivity equal or greater to the models built from curated data is not unexpected. Even if a constructed model has an increased uncertainty arising from having duplicate data points with different activities included, that uncertainty is offset by the probability of having one or more duplicates of compounds in the training set also present in the test set. While this is no advantage if the duplicates have opposing activity classifications or activities differing by several orders of magnitude, being within 1 log  $K_i$  unit of a duplicated structure will be sufficient to bias the external predictivity upwards. In addition, going from 537 to 701 data points in the construction of the model offered a 30% increase in degrees of freedom for fitting the final prediction. Thus, it should not be surprising that, given enough additional data, models

<b><math>\alpha</math>2 adrenergic receptor models</b>			<b>Serotonin reuptake transporter models</b>		
<b>Name</b>	<b>Occurs</b>	<b>Description</b>	<b>Name</b>	<b>Occurs</b>	<b>Description</b>
C3SP3	145	Number of SP3 carbons bound to two other carbons	C2SP3	141	Number of SP2 carbons bound to three other carbons
ATSc2	135	2 <sup>nd</sup> Order Charge Autocorrelation	nAtomP	125	Number of atoms in longest pi-conjugated chain
N2N	131	Count of 2-bond paths between two nitrogen atoms	C1SP3	124	Number of SP3 carbons bound to one other carbon
nBase	114	Number of basic groups	AMR	95	Additive molar refractivity
C2SP3	109	Number of SP3 carbons bound to two other carbons	VCH.5	94	5 <sup>th</sup> order valence chi index
nAromAtom	106	Number of atoms in aromatic systems	MDEC.33	90	Molecular distance edge between all tertiary carbons
C1SP2	105	Number of SP2 carbons bound to one other carbon	C1SP2	90	Number of SP2 carbons bound to one other carbon
C15C	102	Count of 15-bond paths between two carbon atoms	ATSc4	88	4 <sup>th</sup> order charge autocorrelation
Khs.ssNH	101	Number of fragments matching SMARTS [ND2H](-*)-*	C8N	86	Count of 8-bond paths between carbon and nitrogen atoms
ATSc3	99	3 <sup>rd</sup> order charge autocorrelation	ATSc3	82	3 <sup>rd</sup> order charge autocorrelation
WTPT.2	96	Randic index	N8Cl	80	Count of 8-bond paths between nitrogen and chlorine atoms

Alogp2	94	Additive logP	khs.aaaC	78	Number of fragments matching SMARTS [C,c;D3Ho](=*)(=*)-*
VPC.6	87	6 <sup>th</sup> order valence chi index	C3SP3	78	Number of SP3 carbons bound to 3 other carbons
ATSc5	86	5 <sup>th</sup> order charge autocorrelation	khs.dssC	77	Number of fragments matching SMARTS [CD3Ho](=*)(=*)-*
MDEN.22	85	Molecular distance edge between all secondary nitrogens	C7N	76	Count of 7-bond paths between carbon and nitrogen atoms
C6O	85	Count of 6-bond paths between carbon and oxygen atoms	VCH.6	74	6 <sup>th</sup> order valence chi index
N8O	84	Count of 8-bond paths between nitrogen and oxygen atoms	fragC	72	Fragment Complexity Index
S2SP2	83	Number of SP2 carbons connected to 2 other carbons	AlogP	71	Additive logP
Table 4.5: Most commonly occurring descriptors in models (occurring more than twice as often as predicted by random draw)					

can be found which do not lose performance with more noise. A more thorough variant of this computational experiment could be made to test this hypothesis. If the training data were deduplicated randomly (or even by actively attempting to pick the worst reported affinity value) and the chemical structures used for descriptor generation were destandardized after duplicate detection was complete, the model performance metric would not be assured of being the worst possible result, but it would offer a more pessimistic

estimate of the quality of the model generated. The difference between this value and the models generated according the workflow herein described would offer a better estimate of the improvement possible.

## Conclusions

There is sufficient data in ChEMBL to construct QSAR classifier models for a diverse set of drug-like compounds against both alpha-2a and SERT. This data does not translate into a similarly easy path to constructing regression models. Attempts to construct acceptably predictive QSAR regression models of  $\alpha$ 2a models failed over a range of descriptor counts. The small margin between the acceptance threshold and the achieved correlation constants suggests that more aggressive pruning of the data or the addition of novel chemotypes might be sufficient to achieve predictivity.

SERT is known to be promiscuous and this is borne out by the distribution of active and inactive compound in its MDS plot. Conversely,  $\alpha$ 2a is more selective to different chemotypes. Strong evidence of this is seen in the relative number of virtual screening hits found for the respective targets. Additional  $\alpha$ 2a assay data would be useful, as long as it does not introduce more uncertainty into the model. A probable next step would be to look for new  $\alpha$ 2a data in the literature and incorporate it into the extant data set and model. After this, screen ZINC<sup>49</sup> against both models to identify potential leads. There are a large number of compounds in there, so an effective pre-screen would be important. (GDB-17<sup>50</sup> would be interesting to consider with sufficient computing support). Any compound with hits on both  $\alpha$ 2a and SERT, and some subset of compounds falling into the gaps on the MDS plots would be worth seeking experimental assay on, not just to verify the accuracy of the model, but also to determine whether the regions of dual binding affinity are continued through the terra incognita. It would be possible to offer the compounds identified in Figure 4.6 for screening against the  $\alpha$ 2a and SERT receptors by a group such as the PDSP at UNC,

but in view of their relatively low potential for easy functionalization and lack of predicted affinity for SERT, this might well be considered a waste of time and other resources. The scaffolds identified in Figure 4.9, on the other hand, do offer hope for a distinct SAR series. The compounds identified are widely available as screening compounds, and do not offer much novelty, but they are also not a broad class of compounds that have numerous functional substitutions already available in the broader literature or available for commercial purchase. One possible approach to explore the utility of these scaffolds would be to enumerate a collection of virtual analogs (similar to the methods used to enumerate GDB17) and predict the member's affinity at both targets. While these are only computational models, it is not unreasonable to presume that some members of a small virtual library based on these scaffolds would be predicted to show affinity for both targets. If a relatively small number of ligands are predicted to have that affinity, then future exploration is less likely to be profitable, while if a larger fraction are, then investigating custom synthesis may be a more viable option.



## REFERENCES

- (1) Simon, G. E. Social and Economic Burden of Mood Disorders. *Biological Psychiatry* **2003**, *54*, 208–215.
- (2) Lopez-Munoz, F.; Alamo, C. Monoaminergic Neurotransmission: The History of the Discovery of Antidepressants from 1950 until Today. *Current Pharmaceutical Design* **2009**, *15*, 1563–1586.
- (3) Millan, M. Dual- and Triple-Acting Agents for Treating Core and Co-Morbid Symptoms of Major Depression: Novel Concepts, New Drugs. *Neurotherapeutics: the Journal of the American Society for Experimental NeuroTherapeutics* **2009**, *6*, 53–77.
- (4) Andrés, J.; Alcázar, J.; Alonso, J.; Alvarez, R.; Bakker, M.; Biesmans, I.; Cid, J.; Lucas, A.; Drinkenburg, W.; Fernández, J.; Font, L.; Iturrino, L.; Langlois, X.; Lenaerts, I.; Martínez, S.; Megens, A.; Pastor, J.; Pullan, S.; Steckler, T. Tricyclic Isoxazolines: Identification of R226161 as a Potential New Antidepressant That Combines Potent Serotonin Reuptake Inhibition and  $\alpha_2$ -Adrenoceptor Antagonism. *Bioorganic & Medicinal Chemistry* **2007**, *15*, 3649–3660.
- (5) Perović, B.; Jovanović, M.; Miljković, B.; Vezmar, S. Getting the Balance Right: Established and Emerging Therapies for Major Depressive Disorders. *Neuropsychiatric Disease and Treatment* **2010**, *6*, 343–64.
- (6) Kulig, K.; Malawska, B. Estimation of Phospholipophilicity of 1-[3-(arylpiperazin-1-yl)-propyl]-pyrrolidin-2-one Derivatives on Immobilized Artificial Membrane Stationary Phase and Its Correlation with Biological Data. *Biomedical Chromatography* **2006**, *20*, 1129–1135.
- (7) Salt, D.; Maccari, L.; Botta, M.; Ford, M. Variable Selection and Specification of Robust QSAR Models from Multicollinear Data: Arylpiperazinyl Derivatives with Affinity and Selectivity for  $\alpha_2$ -Adrenoceptors. *Journal of Computer-Aided Molecular Design* **2004**, *18*, 495–509.
- (8) Baloch, B.; Jojart, B.; Wagner, Z.; Kovacs, P.; Mate, G.; Gyires, K.; Zadori, Z.; Falkay, G.; Marki, A.; Viskolcz, B. 3D QSAR Models for  $\alpha_2a$ -Adrenoceptor Agonists. *Neurochemistry International* **2007**, *51*, 268276.
- (9) Jabeen, I.; Pleban, K.; Rinner, U.; Chiba, P.; Ecker, G. Structure–Activity Relationships, Ligand Efficiency, and Lipophilic Efficiency Profiles of Benzophenone-Type Inhibitors of the Multidrug Transporter P-Glycoprotein. *Journal of Medicinal Chemistry* **2012**, *55*, 3261–3273.
- (10) Segall, M.; Champness, E.; Leeding, C.; Lilien, R.; Mettu, R.; Stevens, B. Applying Medicinal Chemistry Transformations and Multiparameter Optimization to Guide the Search for High-Quality Leads and Candidates. *Journal of Chemical Information and Modeling* **2011**, *51*, 2967–2976.
- (11) Kaufmann, K.; Dawson, E.; Henry, L.; Field, J.; Blakely, R.; Meiler, J. Structural Determinants of Species-selective Substrate Recognition in Human and Drosophila

Serotonin Transporters Revealed through Computational Docking Studies. *Proteins: Structure, Function, and Bioinformatics* **2009**, *74*, 630–642.

(12) Roman, D.; Walline, C.; Rodriguez, G.; Barker, E. Interactions of Antidepressants with the Serotonin Transporter: A Contemporary Molecular Analysis. *European Journal of Pharmacology* **2003**, *479*, 5353–5363.

(13) Pissurlenkar, R. S.; Dhir, A.; Kessar, S. V.; Kulkarni, S. K.; Coutinho, E. C. An Activity Model for Novel Antidepressants That Interact with the Serotonin Transporter (SERT). *Central Nervous System Agents Medicinal Chemistry* **2011**, *11*, 228–237.

(14) Appell, M.; Dunn, W. J.; Reith, M. E.; Miller, L.; Flippen-Anderson, J. L. An Analysis of the Binding of Cocaine Analogues to the Monoamine Transporters Using Tensor Decomposition 3-D QSAR. *Bioorganic & Medicinal Chemistry* **2002**, *10*, 1197–206.

(15) Kulkarni, S.; Grundt, P.; Kopajtic, T.; Katz, J.; Newman, A. Structure–Activity Relationships at Monoamine Transporters for a Series of N-Substituted 3 $\alpha$ -(Bis[4-Fluorophenyl]methoxy)tropanes: Comparative Molecular Field Analysis, Synthesis, and Pharmacological Evaluation. *Journal of Medicinal Chemistry* **2004**, *47*, 3388–3398.

(16) Kharkar, P.; Reith, M.; Dutta, A. Three-Dimensional Quantitative Structure-Activity Relationship (3D QSAR) and Pharmacophore Elucidation of Tetrahydropyran Derivatives as Serotonin and Norepinephrine Transporter Inhibitors. *Journal of Computer-Aided Molecular Design* **2007**, *22*, 1–17.

(17) Agatonovic-Kustrin, S.; Davies, P.; Turner, J. V. Structure-Activity Relationships for Serotonin Transporter and Dopamine Receptor Selectivity. *Medicinal Chemistry* **2009**, *5*, 271–278.

(18) Han, L.; Wang, Y.; Bryant, S. Developing and Validating Predictive Decision Tree Models from Mining Chemical Structural Fingerprints and High-throughput Screening Data in PubChem. *BMC Bioinformatics* **2008**, *9*, 401.

(19) Pratuangdejkul, J.; Schneider, B.; Launay, J. M.; Kellerman, O.; Manivet, P. Computational Approaches for the Study of Serotonin and Its Membrane Transporter SERT: Implications for Drug Design in Neurological Sciences. *Current Medicinal Chemistry* **2008**, *15*, 3214–3227.

(20) Mente, S.; Gallaschun, R.; Schmidt, A.; Lebel, L.; Vanase-Frawley, M.; Fliri, A. Quantitative Structure-Activity Relationship of Phenoxyphenyl-Methamphetamine Compounds with 5HT<sub>2A</sub>, SERT, and hERG Activities. *Bioorganic & Medicinal Chemistry Letters* **2008**, *18*, 6088–6092.

(21) Walker, T.; Grulke, C.; Pozefsky, D.; Tropsha, A. Chembench: A Cheminformatics Workbench. *Bioinformatics* **2010**, *26*, 3000–3001.

(22) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science* **2008**, *27*, 1337–1345.

- (23) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50*, 1189–1204.
- (24) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29*, 476–488.
- (25) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR & Combinatorial Science* **2007**, *26*, 694–701.
- (26) Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *Journal of Chemical Information and Modeling* **2011**, *51*, 2320–2335.
- (27) Dearden, J.; Cronin, M.; Kaiser, K. How Not to Develop a Quantitative Structure-Activity or Structure-Property Relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research* **2009**, *20*, 241–266.
- (28) Kier, L.; Hall, L.; Murray, W.; Randi, M. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. *Journal of Pharmaceutical Sciences* **1975**, *64*, 1971–1974.
- (29) National Toxicology Program. <https://ntp.niehs.nih.gov/drugmatrix/index.html> (Accessed 24 Oct 2014)
- (30) Gaulton, A.; Bellis, L.; Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research* **2012**, *40*, D1100–7.
- (31) Ruffolo, R.; Bondinell, W.; Hieble, J. alpha- and beta-Adrenoceptors: From the Gene to the Clinic. 2. Structure-Activity Relationships and Therapeutic Applications. *Journal of Medicinal Chemistry* **1995**, *38*, 36813716.
- (32) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, 2014.
- (33) Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, 2013.
- (34) Carhart, R.; Smith, D.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *Journal of Chemical Information and Modeling* **1985**, *25*, 64–73.
- (35) O’Boyle, N. Molecular Graph-ics with Pybel. <http://baoilleach.blogspot.com/2008/10/molecular-graph-ics-with-pybel.html> (Accessed 25 Oct 2014).
- (36) Rossum, G. van; Drake, F. L. *Python Reference Manual*; PythonLabs: Virginia, 2001.
- (37) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics and Function Using NetworkX, 2008.

- (38) O'Boyle, N.; Morley, C.; Hutchison, G. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chemistry Central Journal* **2008**, *2*, 5.
- (39) Scipy.org. <http://scipy.org> (Accessed on 25 Oct 2014).
- (40) Guha, R. CDK Descriptor Calculator, <http://www.rguha.net/code/java/cdkdesc.html> (Accessed on 25 Oct 2014).
- (41) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Current Pharmaceutical Design* **2007**, *13*, 3494–504.
- (42) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- (43) Golbraikh, A.; Tropsha, A. Beware of  $q^2$ ! *Journal of Molecular Graphics & Modelling* **2002**, *20*, 269–276.
- (44) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Koetter, T.; Meinel, T.; Ohl, P.; Seib, C.; Tiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. Springer: New York, **2007**.
- (45) Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **1912**, *11*, 37–50.
- (46) Roth, B. NIMH Psychoactive Drug Screening Program. <http://pdspdb.unc.edu/pdspWeb> (Accessed on 25 Oct 2014).
- (47) Summary of Capabilities. <https://pharmacy.unc.edu/research/centers/center-for-integrative-chemical-biology-and-drug-discovery/summary-of-capabilities>. (Accessed on 25 Oct 2014).
- (48) Enamine Ltd. <http://www.enamine.net/index.php> (Accessed on 25 Oct 2014).
- (49) Irwin, J.; Sterling, T.; Mysinger, M.; Bolstad, E.; Coleman, R. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modelling* **2012**, *52*, 1757–1768.
- (50) Ruddigkeit, L.; Deursen, R.; Blum, L.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–75.
- (51) Baell, J.B.; Holloway, G.A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry* **2010**, *53*, 2719–2740.

## Chapter 5: Conclusions

*“It is not incumbent on you to complete the work, but neither are you at liberty to desist from it”*

Pirkei Avot 2:21

### **Towards rational data curation of biochemical affinity data**

It is clear from both the extant literature and from the experiences contained in this work that careful evaluation and curation of available data is required for the construction of predictive QSAR models. In Chapter Two, we have shown that no single database contains the correct structure for even the most common small molecule drugs on the market today. In Chapter Three, we explored the errors found in the representations of biological activity within one of the broadest, freely-available repositories. In Chapter Four, we utilized a curation workflow in order to develop QSAR models for potential bi-targeted therapeutic agents.

The curation work presented here leaves us with a paradox. We have shown that automated methods alone cannot curate, because they will lack the experience and intuition of a human chemist to understand from whence a problem arises. But humans alone also cannot effectively curate, because they tend to work slowly and to introduce errors from manual handling of data. The most likely way forward is semi-automated curation where a computer identifies likely problems in data (and a random sampling of all other data) that are then dealt with and/or verified by a human curator.

While the use of consensus methods to detect erroneous data is attractive, no doubt in part because of its apparent simplicity, its applicability is inherently limited because over

90% of the compounds reported in the medicinal chemistry literature appear exactly once. Crosschecking structures between the peer-reviewed and the patent literature may become more rational with the recent opening of access to Surechem<sup>1</sup>, but given that Linceptor<sup>2</sup> and ChEMBL<sup>3</sup> have an overlap of only 3%, this is not likely to significantly increase the number of independent reports of novel chemical matter. Additionally, with the current practices of pharmaceutical companies and patent practitioners as exemplified by the Lilly Viagra patent case in Canada<sup>4</sup>, identifying the corresponding compounds between a journal report and the patent application will only get more difficult. This is not intended to denigrate the utility of multiple database entries in identifying transcription errors, but these are only a portion of the errors present in the literature. These errors will also remain relatively tractable as long as there are multiple independent data aggregators in cheminformatics. Of course, this requires data consumers to be aware of the provenance of their data. Knowing that Wikipedia sources most of its chemical structure data directly from Chemspider<sup>5</sup>, or that Pubchem and ChEMBL include substantial portions of each other's data in their own database greatly simplifies the process of eliminating erroneous structures from a dataset. But as the number of freely-reusable chemical activity databases grows, the risk that information is silently transferred from one to the next instead of being independently transcribed from the literature only increases. If information is incorrect in the original publication, there is little that can be done with it except set it aside and hope for a correction in the coming months. When the error arises further along in the data compilation and distribution process, it can be identified and corrected as long as backtracking to the original publication is possible.

The optimal solution for the problem of data accuracy in cheminformatics is, as suggested earlier, the adoption of the MIABE<sup>6</sup> standards by the research publishing community. Requiring the direct deposit of structural and activity data in appropriate repositories is already a condition of publication in molecular biology, protein

crystallography, and bioinformatics. With essentially all data and manuscripts in the medicinal chemistry literature prepared electronically and most journals preferentially publishing in electronic form, inserting steps which only serve to prevent data from being efficiently and accurately extracted makes sense only if the authors and editors are attempting to limit the reproduction of the data being reported.

While the scarcity of replication impacts the ability to determine correct chemical structures, it even more strongly impacts the availability of pharmacological profiles associated with those structures. Accordingly, we have put forward a workflow for the curation of biological data to be used in QSAR modeling. Compound deduplication by minimizing the cumulative systematic error appears to be the most viable approach. While this is not a complicated concept or counter-intuitive in any significant way, it has never been formally proposed in the literature, even when best practices for QSAR modeling have been discussed. Indeed, it appears to have never even reached the status of oral folklore within the community, either as a positive or negative approach.

We have also developed models for  $\alpha_2$  adrenergic receptor and SERT binding affinity with high internal and external predictivity. While they have not yet undergone experimental validation, they have suggested the availability of new chemical space for exploration of  $\alpha_2$  binding affinity. In addition, virtual screening of the models against commercial catalog has identified compounds known to possess binding affinity at the receptors but not included in the training data.

### **Data density and the inclusion of data from multiple sources in a single modeling set**

The broad error band associated with experimental values published in the literature presents a special challenge to those who seek to use that data as the basis of molecular models. Given that half of the binding targets from the primary literature appearing in

ChEMBL have three or fewer ligands associated with them, there are many cases where insufficient data exists for any model to be constructed. Even if a large scale HTS diversity campaign has been completed for the target, such as DrugMatrix<sup>7</sup>, the density of data points will be relatively low. If all that is necessary is to predict whether unknown compounds will bind at micromolar concentrations, this is probably sufficient, but trying to prioritize possible leads from virtual screening into nanomolar concentrations may well prove futile.

There has been a significant amount of discussion regarding whether it is desirable, or even possible to combine data from different sources to construct increasingly general models for activity at a single target instead of limiting efforts to more local models which rely on a single SAR series of compounds as their primary basis. Consensus best practices with the QSAR modeling community would suggest that five-fold external cross-validation should be utilized in model construction to minimize the risk of overfitting<sup>8</sup>. This implies that data for at least ten molecules are required before starting model building, or two molecules for each fold (one active and one inactive in a classifier model, or two points to define a line in a regression model). Modeling lore has many rules of thumb for how many descriptors can safely be included in a model given the number of molecules in the dataset or the size of the smaller class in a classifier model. These range from a ratio of one descriptor for every five molecules, to one of one descriptor for every twenty molecules. Similarly, received wisdom suggests that no classifier model should have a size imbalance of more than 3:1 or 4:1 between the two classes (this can be resolved by selectively down-sampling the larger set, of course)<sup>9</sup>. In the case of regression models, skewed distributions of compounds in general are disfavored; a relatively uniform distribution of activities over the entire range is preferable. If these rules of thumb are used loosely, a two descriptor model can be made on the smallest possible data set (ten molecules evenly distributed). However, this is still a significant barrier for using data collected in ChEMBL. In Chapter Three, it was noted that the median number of distinct molecules in a single paper in ChEMBL was 14, with an inter-



quartile distance of 19. This suggests that only about 60% of the papers in ChEMBL could be the basis of some sort of QSAR model. But the conditions imposed are the loosest possible. If more rigorous conditions are imposed, such as requiring 3 or 4 compounds in each class of a fold, more than 33 compounds will be required, excluding 75% of all publications in ChEMBL. To reach a point where a five descriptor model is reasonably usable if we require ten molecules per descriptor in the smaller class, we will need 100 molecules total (which fewer than 200 papers have). While many small datasets may yield well-behaved models, larger consensus models require more molecules than can be reliably found in the typical report in the medicinal chemistry literature.

The risks involved in combining multiple datasets may, however, be overestimated. While it appears that any given affinity measurement is only accurate to 1 pK<sub>i</sub> unit, only part of that quantity is due to truly random error. By ensuring that each data source incorporated in the training set is large enough to have at least one comparable value in each fold and limiting the number of different sources used to cover the chemical space being modeled, the systemic error is reduced, if not minimized for practical purposes. Being mindful of the inherent uncertainty in binding affinity assays also highlights the importance of not expecting too much from one's models. If the uncertainty in a binding affinity is 1 pK<sub>i</sub> unit, there is little rationale in expecting the model to be accurate to significantly less than that. Being able to distinguish between compounds that are micromolar, nanomolar, and picomolar is a practical goal that can be accomplished with classifier models, but to demand the distinction between 1 nanomolar and 10 nanomolar is hardly rational when those values cannot be experimentally determined.

### **Further work and directions**

As previously discussed, quantifying the improvement of the proposed workflow and demonstrating its utility for general use remains to be done. The inability of the trial

attempted in Chapter Four to make a noticeable change in  $R^2$  when no deduplication was attempted was hampered by the increased degrees of freedom afforded by a 30% increase in the number of data points. In order to properly test the performance of this workflow, a series of computational experiments will be necessary. Each trial would ideally be based on an already-built model with experimentally-validated results at a different target. For each target, the uncurated data set used for training and testing of the model would either be re-extracted from the origination database or otherwise reassembled. In one case, the model would be rebuilt using the workflow described herein and then used for virtual screening of the originally screened library. In a second case, the deduplication process would be either random or designed to use the affinity value with the least congruence to other values in the dataset (favor one-off compounds and pick values from the smallest series possible).

Similarly, instead of standardizing chemical structures to a canonical form, resonance structures and atom-typing would be randomized. This step is particularly important if all the chemical structures came from a database such as ChEMBL that does standardize structures as part of their own workflow. Finally, it may be possible to interchange the affinity values for a small percentage of highly similar compounds in the data set to simulate the effects of errors in transcription of structures. Once this “poor” dataset is assembled, it would likewise be used to build QSAR models and screen a virtual library for novel chemical matter.

This approach would provide three numerical benchmarks for each dataset:  $Q^2$ ,  $R^2$  (or internal and external CCR), and the number of hits retrieved from the virtual screening library. These could each be compared to the original model that was constructed for a more complete idea of how each step of our standard workflow is affected by data quality. Similarly, the actual structures of the hits returned can be compared to assess the impact on predictions based on models from data sets of differing quality. This design would also allow

for multiple tests to be conducted at targets of differing promiscuity and biological nature (membrane-bound vs. free, receptor vs. enzyme vs. transporter).

Another area for future work would be to consider the average distances between members of a SAR series compared to drug-like molecules or ChEMBL as a whole. When a full chemical activity dataset is projected into a two-dimensional plot as seen in chapter four, compounds present as members of individual SAR series will tend to cluster together. Conversely, results from a diversity-oriented HTS screen will scatter more widely and not partition themselves into a small portion of the available space. This is a semi-graphical approach to work done by Denis Fourches, who recently used a clustering of CYP450 substrates by chemical structure to identify compounds with incorrect structures. Compounds which are not substrates and are sorted into a cluster where the compounds are predominantly metabolized by the specific isoform (or vice versa) may be true activity cliffs, but they also at an increased risk for having errors associated with them, either in their chemical structure or in the reported assay results. By examining specific instances where members of a series either form an activity cliff with another member of the series, or are anomalously distant from all the other members of the series, it will be easier to diagnose problems with a data set that need to be addressed by further research by a human chemist. For example, activity cliffs do exist in real data sets (often representing a shift in binding mode, or when a substituent has become large enough that a binding pocket can no longer accommodate it), but two compounds of similar structure and similar binding affinity would also appear to be an activity cliff if the units for the affinity of one of the pair to a common target were incorrectly recorded (*e.g.* micromolar instead of nanomolar or vice versa). On the other hand, a compound which is nominally within a SAR series, and yet is separated from all other members of the series in chemical space may well have an error in its structure. This is not absolute, as we see in figure 6 of chapter 4, but it again provides a means to prioritize information in the dataset needing verification.

To reiterate, fully automated and completely manual curation schemes are both inherently flawed. Computers can rapidly and accurately copy and compare chemical structures and biological data, but their capacity to identify why a value or structure is problematic is limited to patterns that have already been identified and coded. Human chemists can apply pattern recognition and imagination to identify which data points are incorrect and the reason for the inaccuracy, but they work relatively slowly and are prone to errors when manually transcribing information. In the absence of canonical information coming directly from the originator of the data, errors will need to be identified by combining the accuracy and computational speed of a computer with the intuition and creativity of a human mind.

## REFERENCES

- (1) Overington, J. P. SureChEMBL - Chemical Structure Information in Patents. <http://chembl.blogspot.com/2013/12/surechembl-chemical-structure.html> (Accessed on 25 Oct 2014).
- (2) Linceptor Database. <http://liceptor.com/products/databases/liceptordatabase.html> (Accessed on 24 October 2014)
- (3) Bento, A.; Gaulton, A.; Hersey, A.; Bellis, L.; Chambers, J.; Davies, M.; Krüger, F.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Research* **2014**, *42*, D1083–D1090.
- (4) Supreme Court of Canada. *Teva Canada Ltd. v. Pfizer Canada Inc.*; [2012] 3 R.C.S.; 2012.
- (5) Williams, A.; Tkachenko, V. The Royal Society of Chemistry and the Delivery of Chemistry Data Repositories for the Community. *Journal of Computer-Aided Molecular Design* **2014**.
- (6) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; Glen, R.; Grigorov, M.; Hammond-Kosack, K.; Harland, L.; Hopkins, A.; Larminie, C.; Lynch, N.; Mann, R.; Murray-Rust, P.; Piparo, E.; Southan, C.; Steinbeck, C.; Wishart, D.; Hermjakob, H.; Overington, J.; Thornton, J. Minimum Information about a Bioactive Entity (MIABE). *Nature Reviews. Drug Discovery* **2011**, *10*, 661–669.
- (7) National Toxicological Program. DrugMatrix. <https://ntp.niehs.nih.gov/drugmatrix/index.html> (Accessed 25 Oct 2014).
- (8) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29*, 476–488.
- (9) Dearden, J.; Cronin, M.; Kaiser, K. How Not to Develop a Quantitative Structure-Activity or Structure-Property Relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research* **2009**, *20*, 241–266.

## APPENDIX 1: NAMES OF DRUGS FOR INTERNET STRUCTURE SEARCH

Atorvastatin  
Clopidogrel  
Amlodipine  
Olanzapine  
Valsartan  
Risperidone  
Venlafaxine  
Pantoprazole  
Montelukast  
Quetiapine  
Lansoprazole  
Losartan  
Alendronate  
Pioglitazone  
Simvastatin  
Rabeprazole  
Imatinib  
Zolpidem  
Donepezil  
Donepezil  
Cetirizine  
Irbesartan  
Irbesartan  
Docetaxel  
Oxaliplatin  
Sertraline  
Oseltamivir  
Celecoxib  
Topiramate  
Bupropion  
Aripiprazole  
Lamotrigine

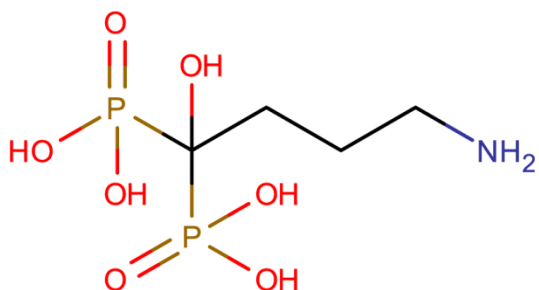
Metoprolol  
Candesartan  
Sildenafil  
Telmisartan  
Leuprolide  
Fenofibrate  
Ondansetron  
Valaciclovir  
Levofloxacin  
Anastrozole  
Tacrolimus  
Mycophenolate mofetil  
Latanoprost  
Carvedilol  
Gemcitabine  
Omeprazole  
Duloxetine  
Sumatriptan  
Fentanyl  
Budesonide  
Zoledronate  
Ramipril  
Bicalutamide  
Raloxifene  
Tamsulosin  
Pregabalin  
Paroxetine  
Lopinavir  
Tolterodine  
Tamsulosin  
Goserelin  
Levofloxacin  
Drospirenone  
Terbinafine  
Piperacillin

Tadalafil  
Levetiracetam  
Atazanavir  
Methylphenidate  
Ciclosporin  
Somatostatin  
Irinotecan  
Fexofenadine  
Amphetamine  
Nifedipine  
Moxifloxacin  
Meloxicam  
Clarithromycin  
Sevoflurane  
Efavirenz  
Linezolid  
Capecitabine  
Ziprasidone  
Ciprofloxacin  
Modafinil  
Fluvastatin  
Desloratadine  
Letrozole  
Oxcarbazepine  
Bosentan  
Temozolomide  
Dorzolamide  
Diclofenac  
Tenofovir  
Pramipexole  
Memantine  
Ramipril  
Azithromycin  
Cefdinir  
Finasteride

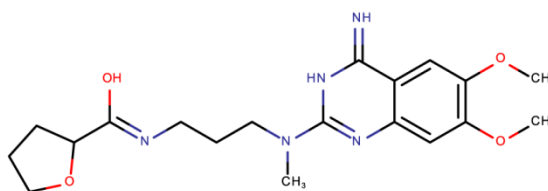


Pemetrexed  
Meropenem  
Atomoxetine  
Fentanyl  
Glimepiride  
Lidocaine  
Eszopiclone  
Paclitaxel  
Tegaserod  
Levalbuterol  
Orlistat  
Enalapril  
Salmeterol  
Doxazosin  
Levothyroxine  
Famotidine  
Caspofungin  
Rivastigmine  
Voriconazole  
Amlodipine  
Niacin  
Gabapentin  
Ropinirole  
Voglibose  
Metformin  
Bisoprolol  
Alfuzosin  
Fluconazole  
Thalidomide  
Ranitidine  
Loratadine

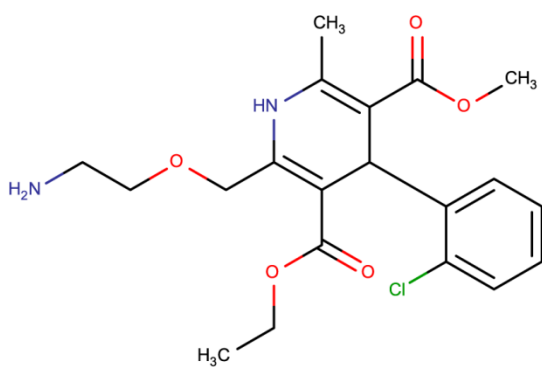
APPENDIX 2: GOLD LIST FINAL CONSENSUS DRUG STRUCTURES



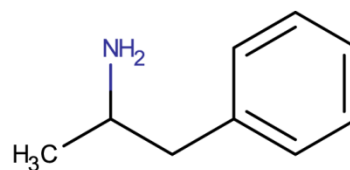
Alendronate



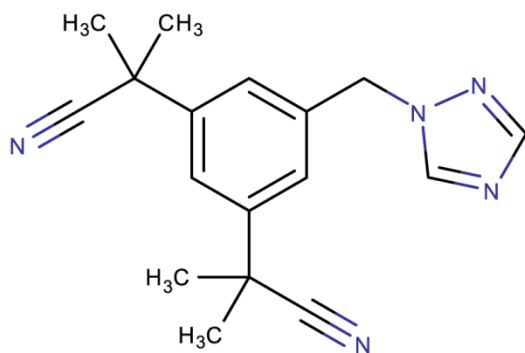
Alfuzosin



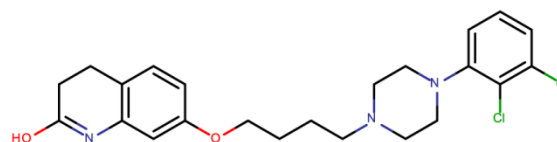
Amoldipine



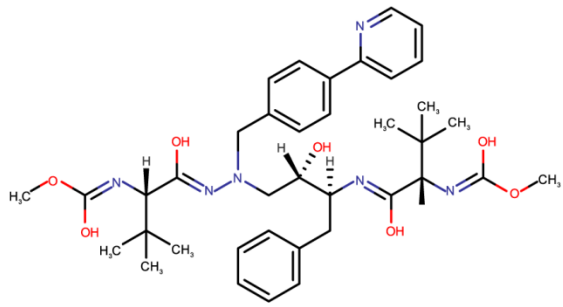
Amphetamine



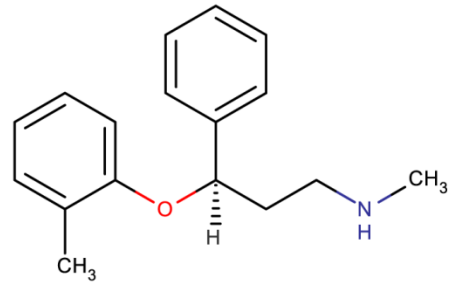
Anastrozole



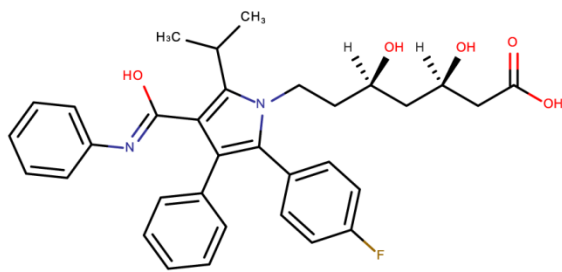
Aripiprazole



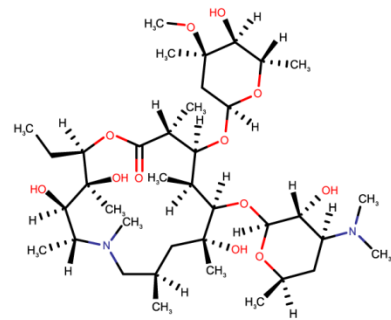
Atazanavir



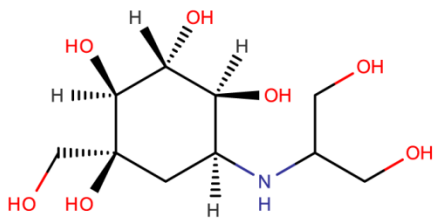
Atomoxetine



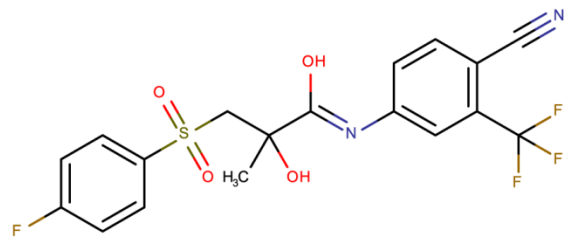
Atorvastatin



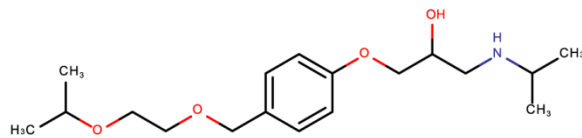
Azithromycin



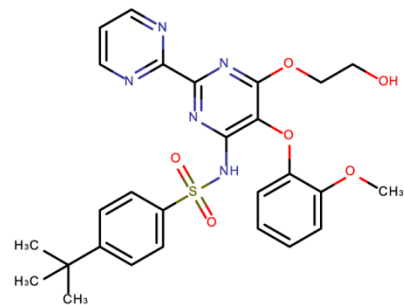
Basen



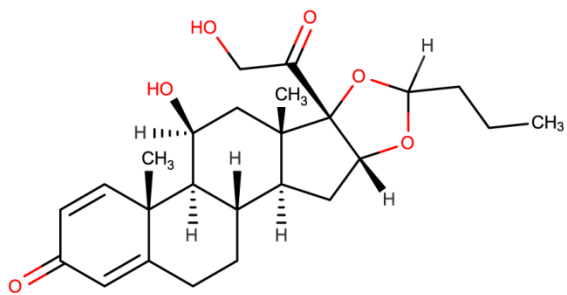
Bicalutamide



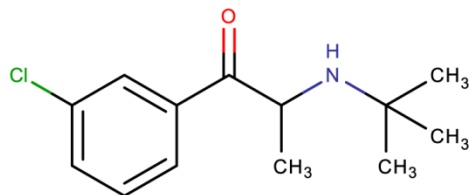
Bisoprolol



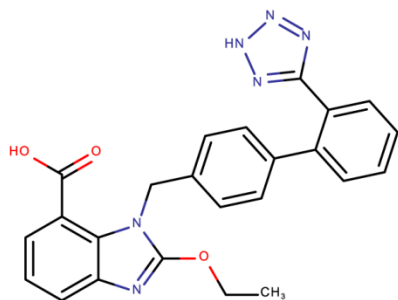
Bosentan



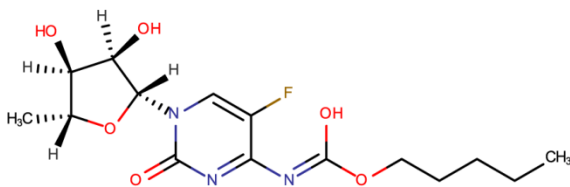
Budesonide



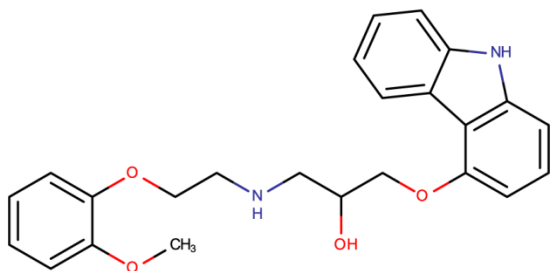
Bupropion



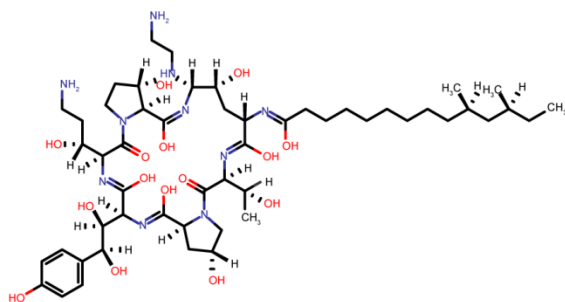
Candesartan



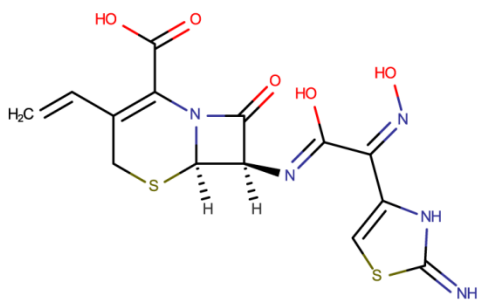
Capecitabine



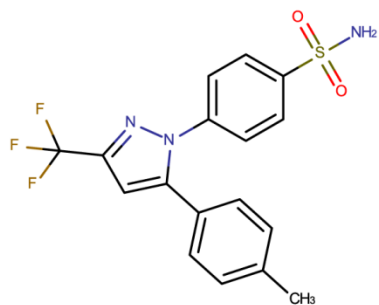
Carvedilol



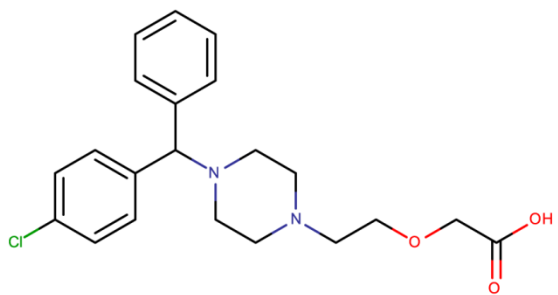
Caspofungin



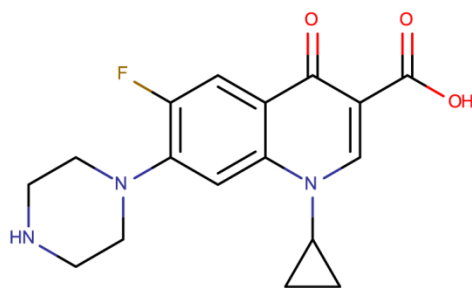
Cefdinir



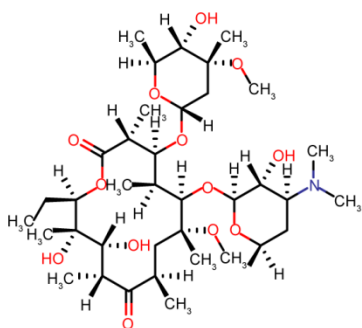
Celecoxib



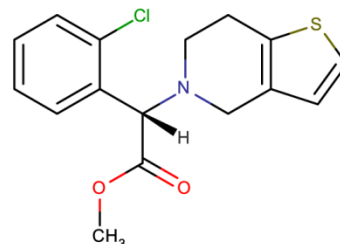
Cetirizine



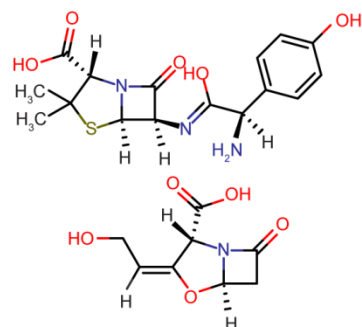
Ciprofloxacin



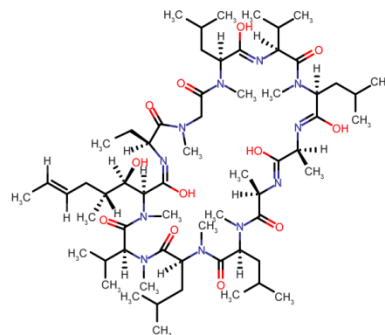
Clarithromycin



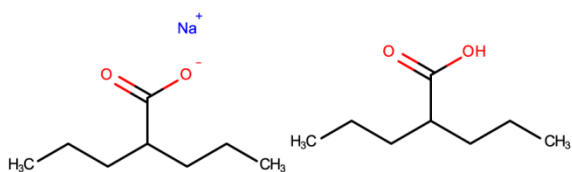
Clopidogrel



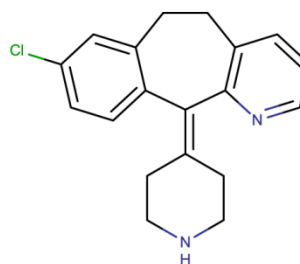
Co-amoxiclav



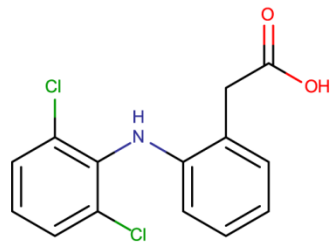
Cyclosporin



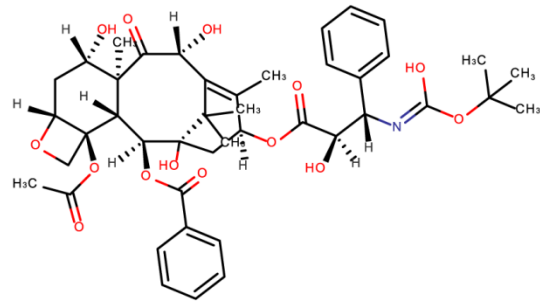
Depakote



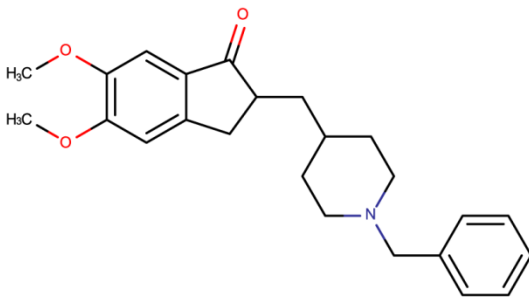
Desloratadine



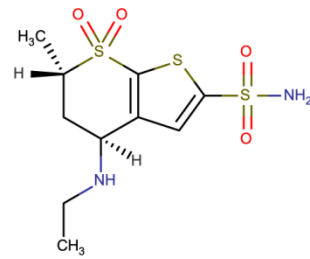
Diclofenac



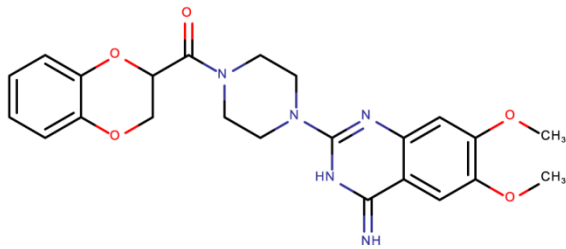
Docetaxel



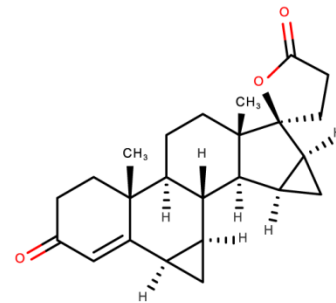
Donepezil



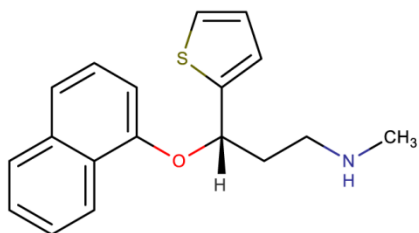
Dorzolamide



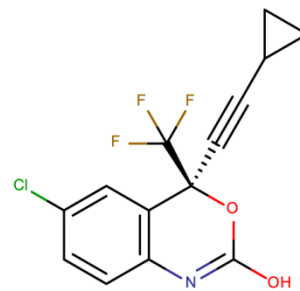
Doxazosin



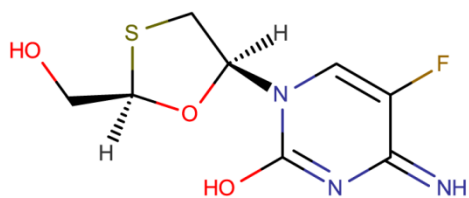
Drospirenone



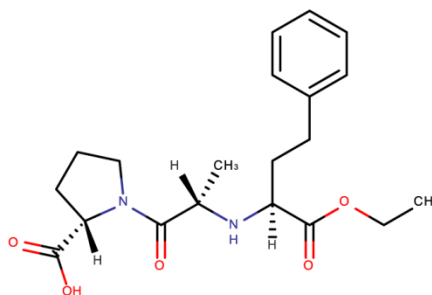
Duloxetine



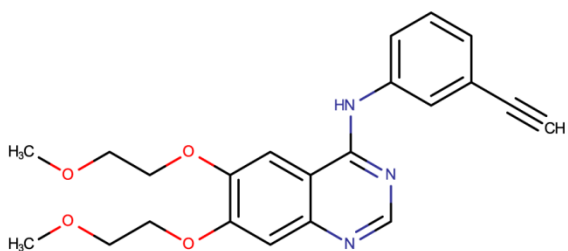
Efavirenz



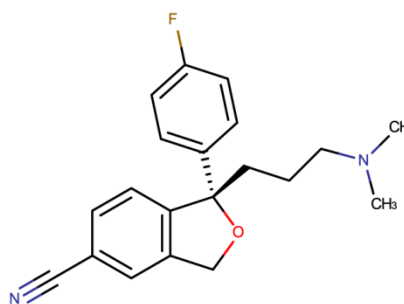
Emtricitabine



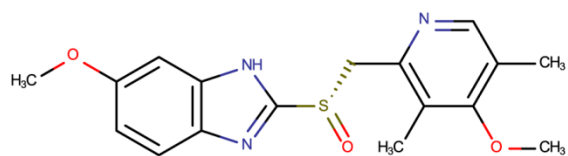
Enalapril



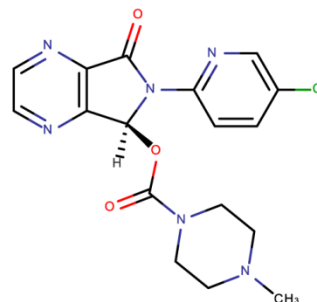
Erlotinib



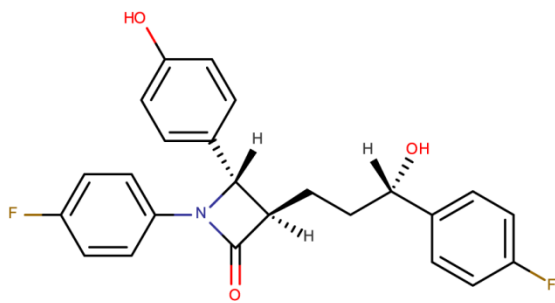
Escitalopram



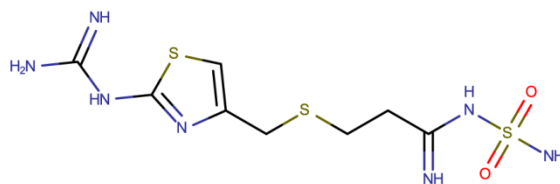
Esomeprazole



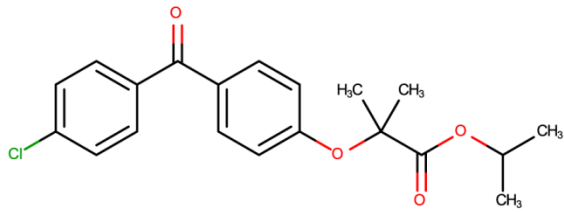
Eszopiclone



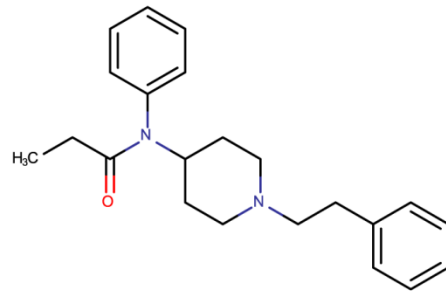
Ezetimibe



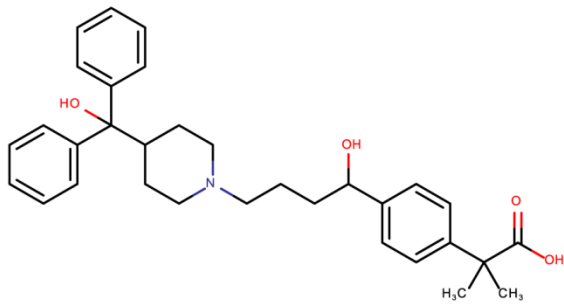
Famotidine



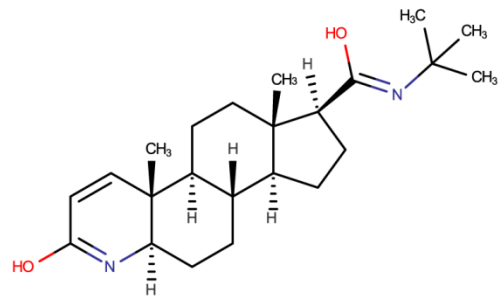
Fenofibrate



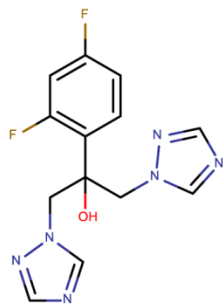
Fentanyl



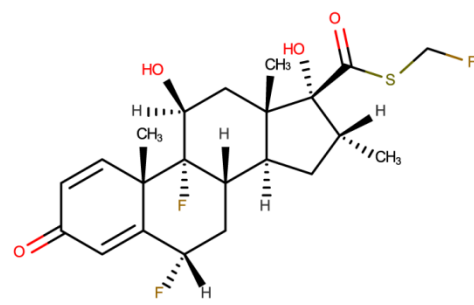
Fexofenadine



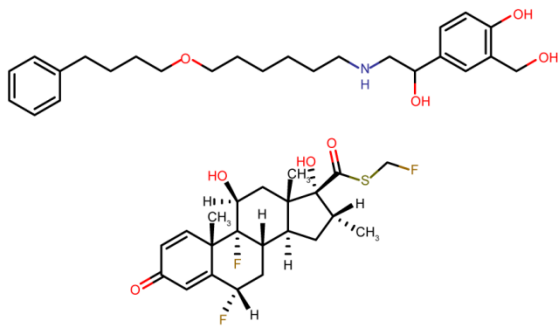
Finasteride



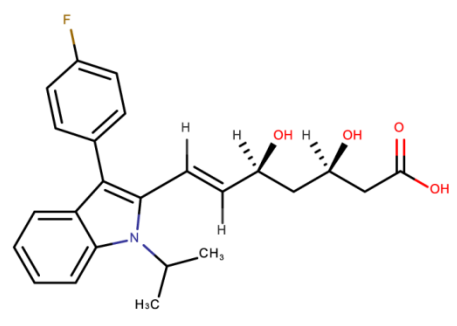
Fluconazole



Fluticasone

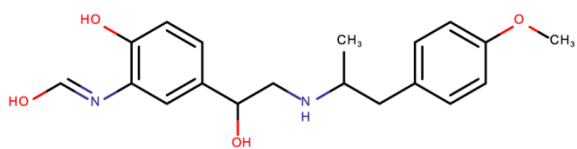


Fluticasone/Salmeterol

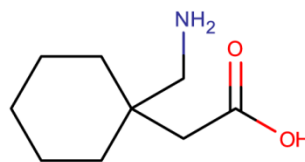


Fluvastatin

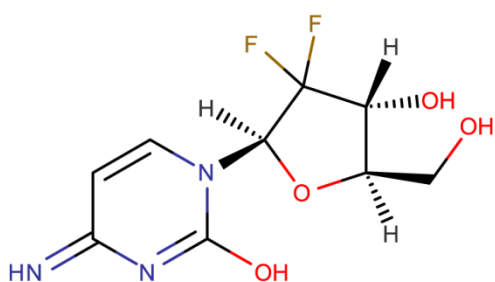




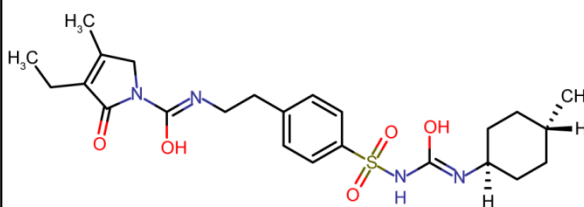
Formoterol



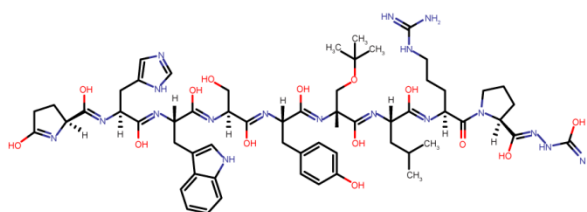
Gabapentin



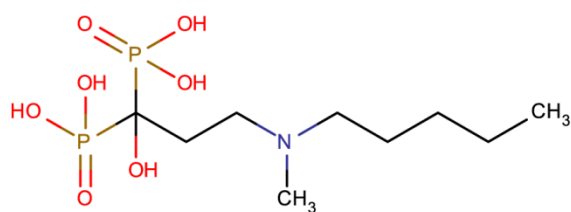
Gemcitabine



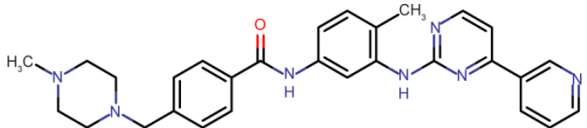
Glimepiride



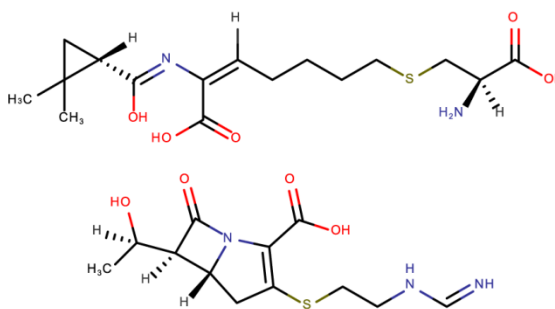
Goserelin



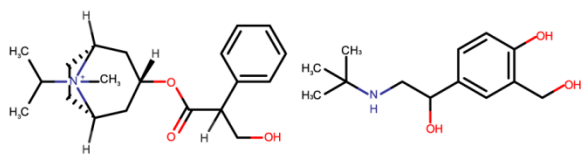
Ibandronate



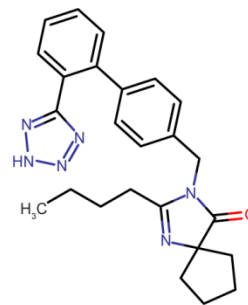
Imatinib



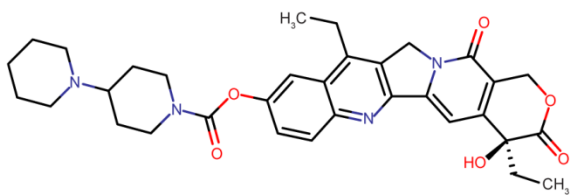
Imipenem/Cilastatin



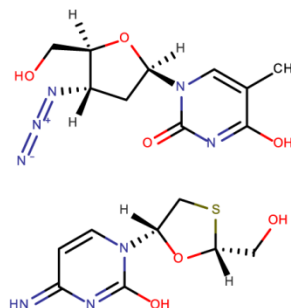
Ipratropium/Salbutamol



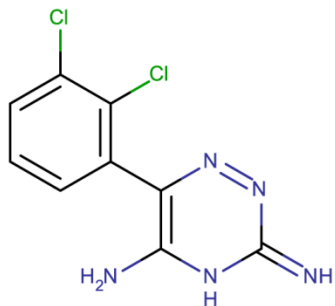
Irbesartan



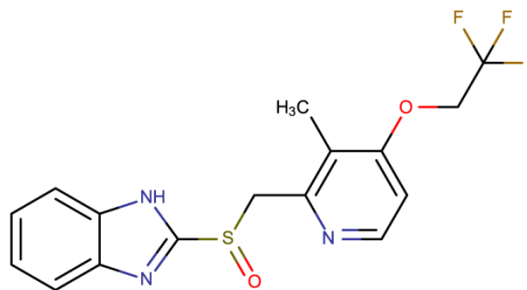
Irinotecan



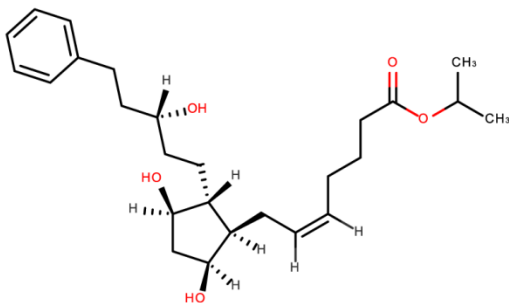
Lamivudine/Zidovudine



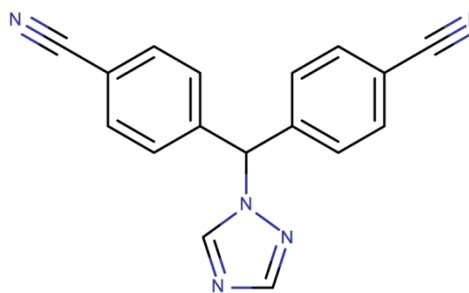
Lamotrigine



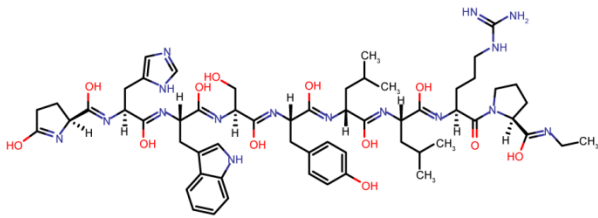
Lansoprazole



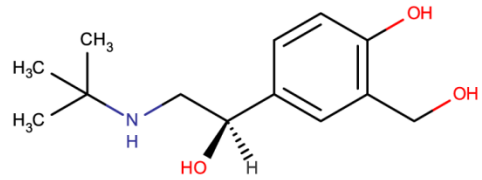
Latanoprost



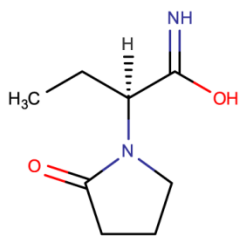
Letrozole



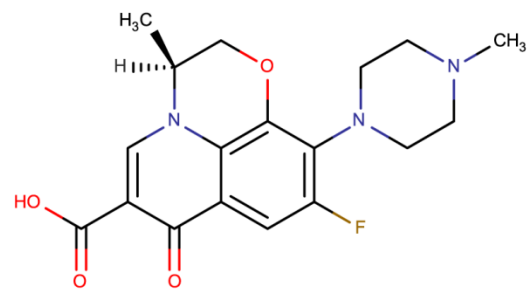
Leuprolide



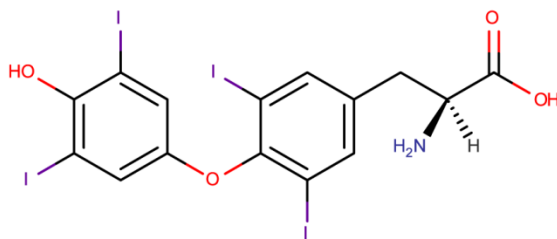
Levalbuterol



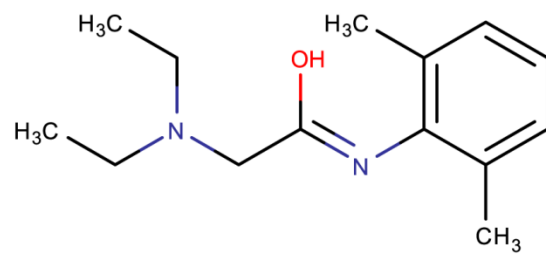
Levetiracetam



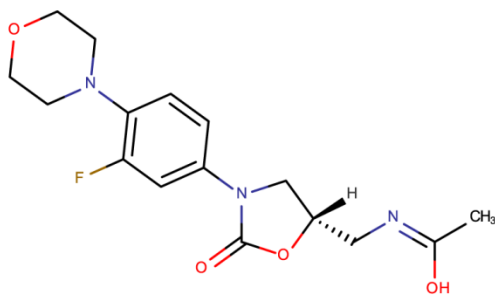
Levofloxacin



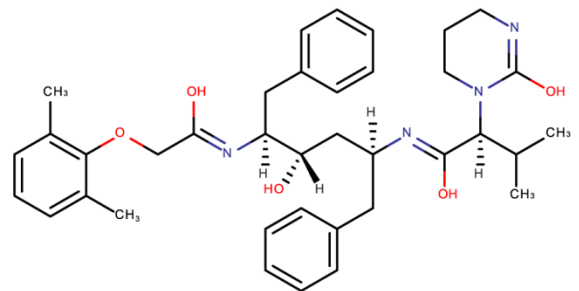
Levothyroxine



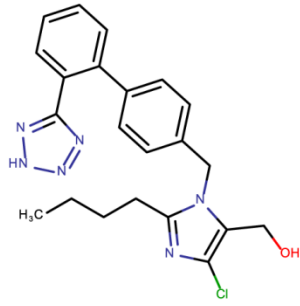
Lidocaine



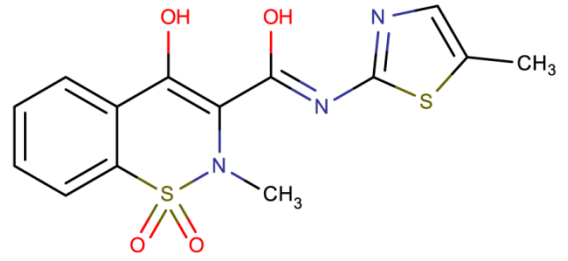
Linezolid



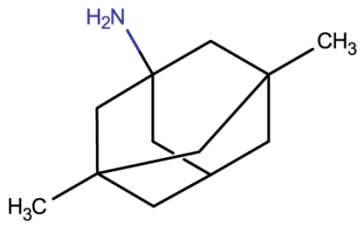
Lopinavir



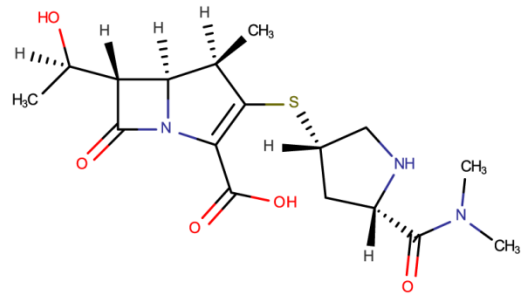
Losartan



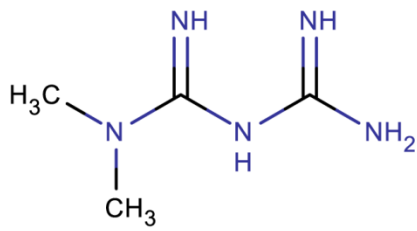
Meloxicam



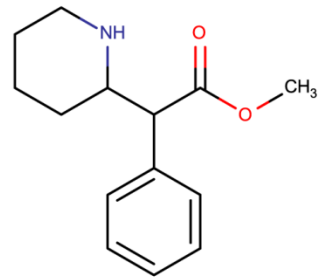
Memantine



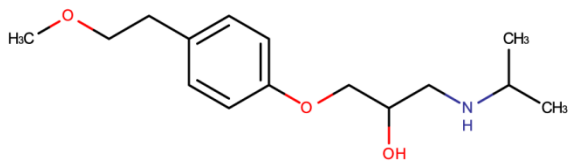
Meropenem



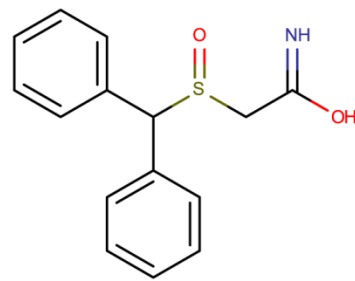
Metformin



Methylphenidate

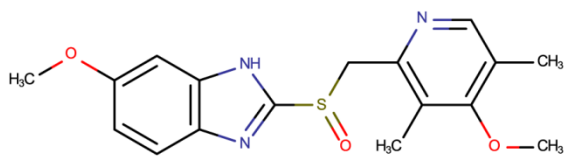


Metoprolol

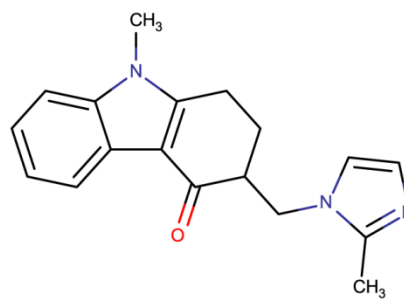


Modafinil



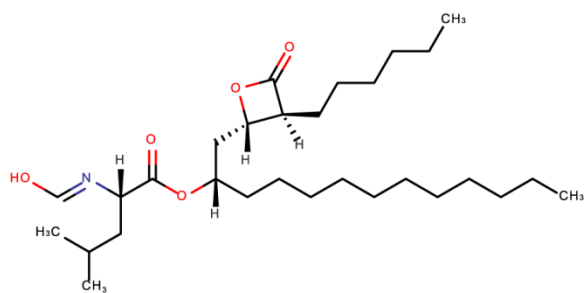


Omeprazole

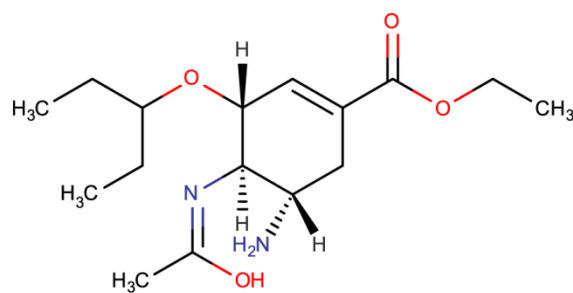


Ondansetron

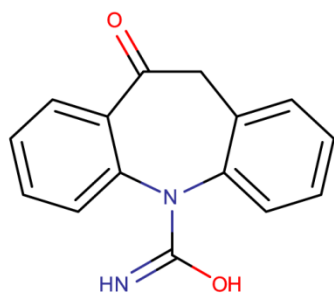
92



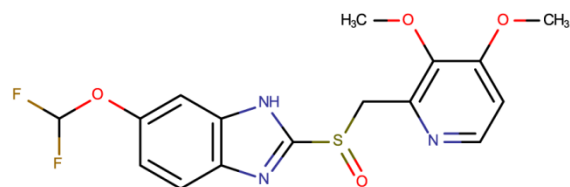
Orlistat



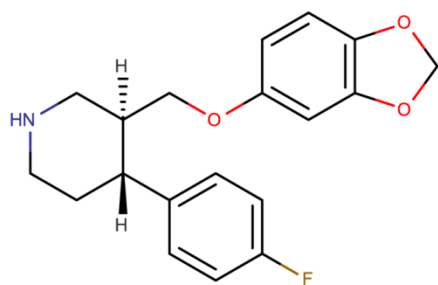
Oseltamvir



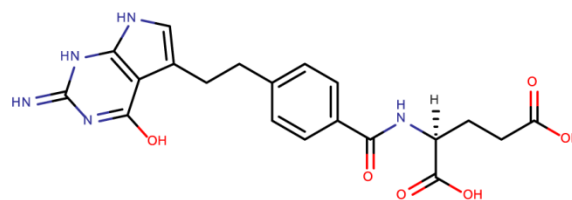
Oxcarbazepine



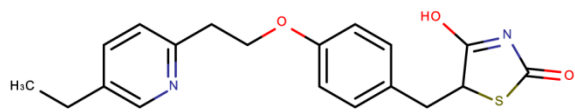
Pantoprazole



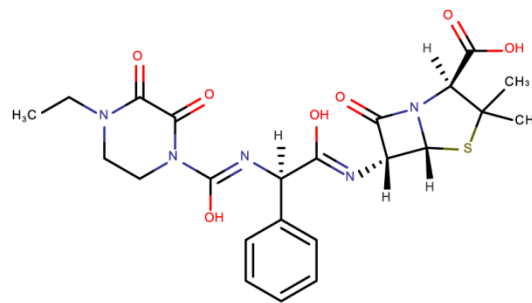
Paroxetine



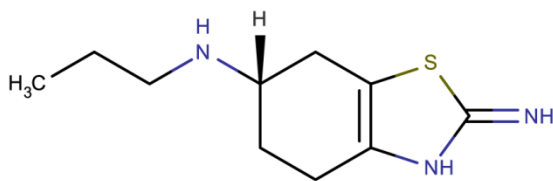
Pemetrexed



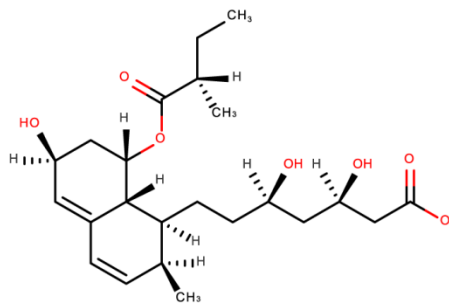
Pioglitazone



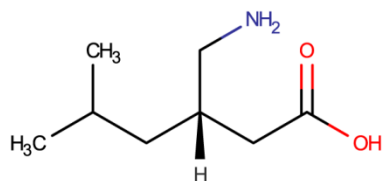
Piperacillin



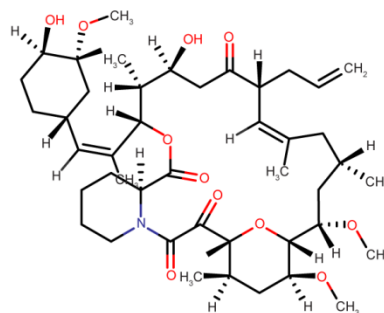
Pramipexole



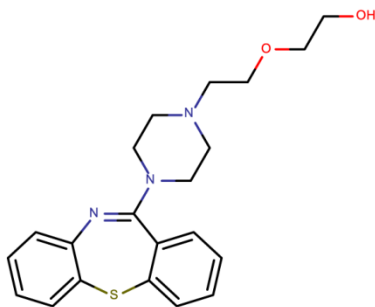
Pravastatin



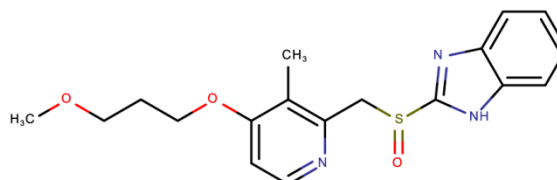
Pregabalin



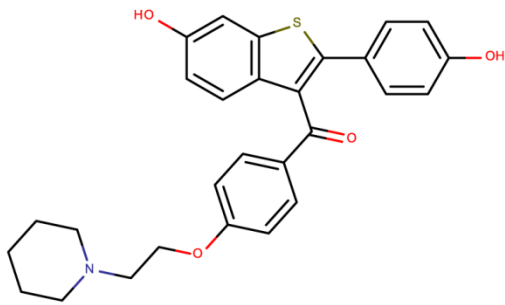
Prograf



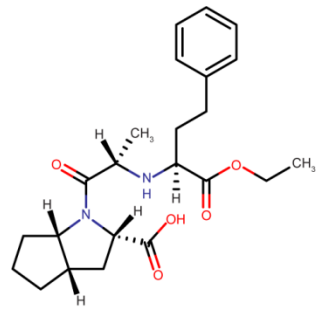
Quetiapine



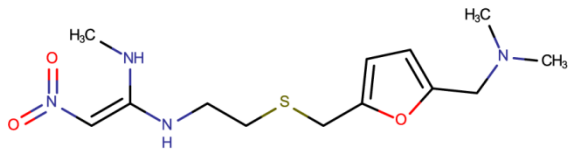
Rabeprazole



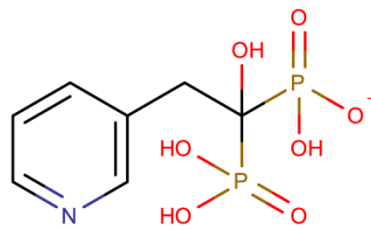
Raloxifene



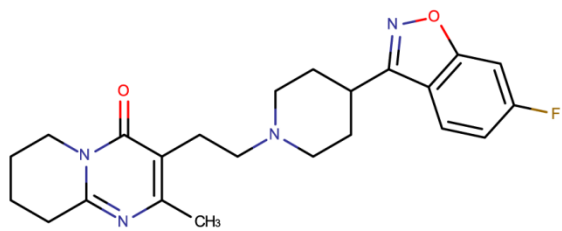
Ramipril



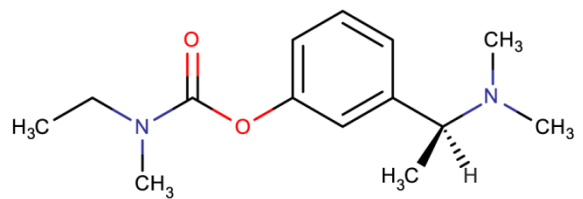
Ranitidine



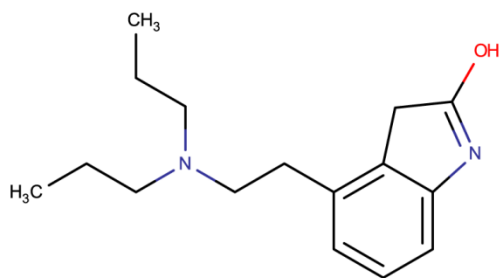
Risedronate



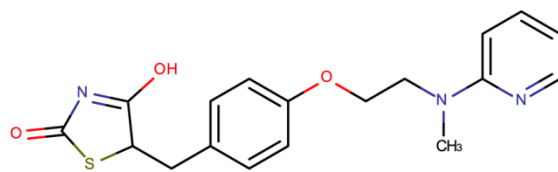
Risperidone



Rivastigmine

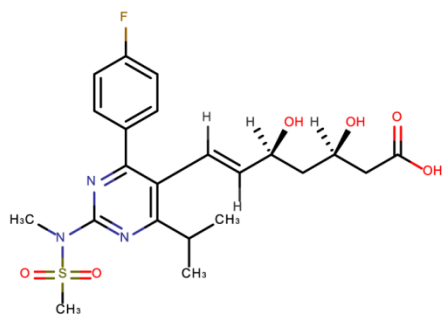


Ropinirole

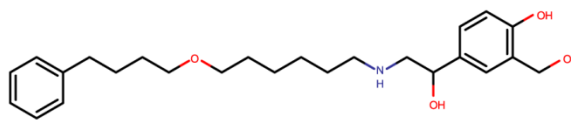


Rosiglitazone

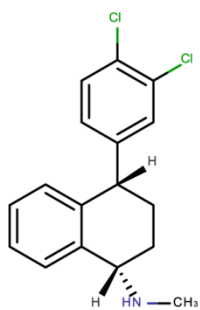




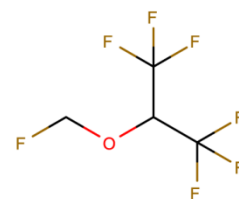
Rosuvastatin



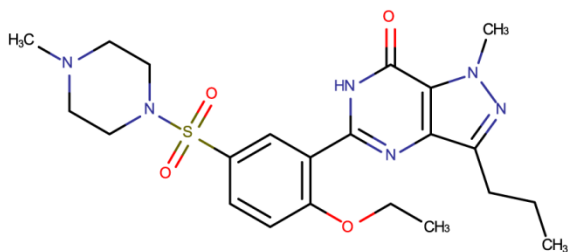
Salmeterol



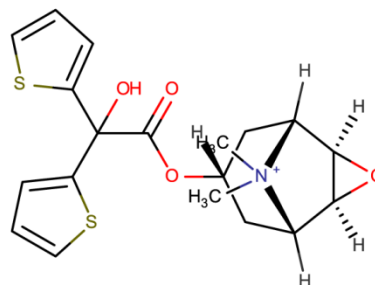
Sertraline



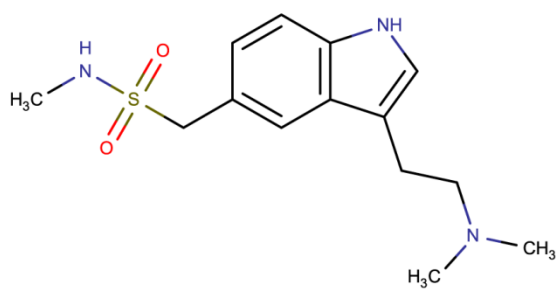
Sevoflurane



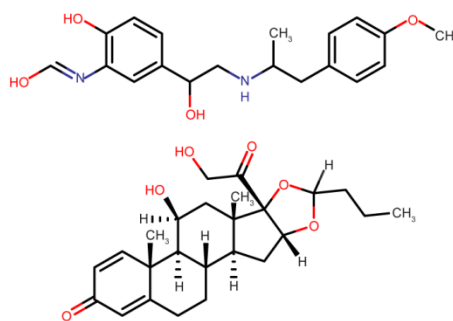
Sildenafil



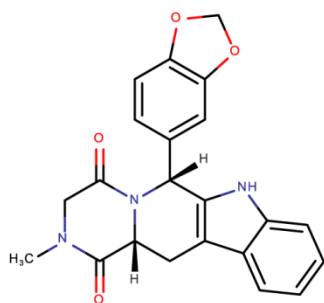
Spiriva



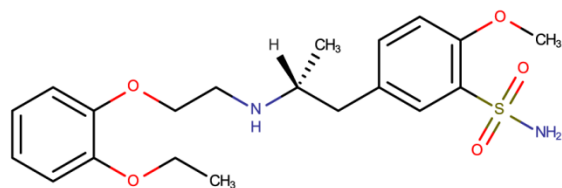
Sumatriptan



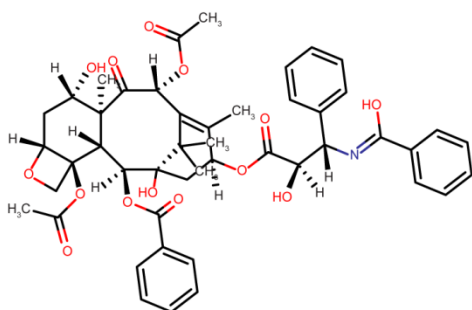
Symbicort



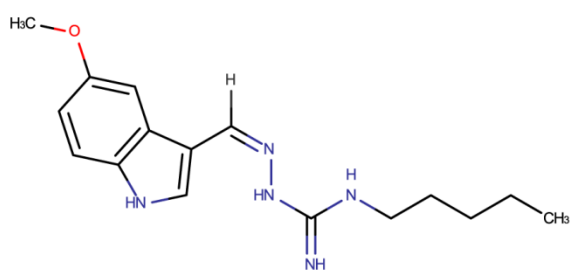
Tadalafil



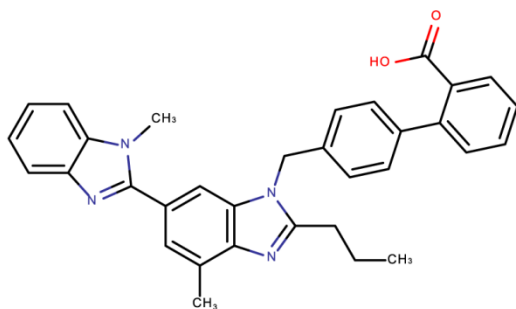
Tamsulosin



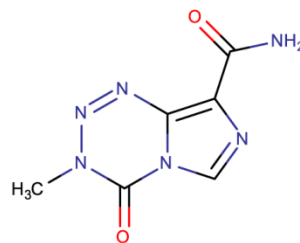
Taxol



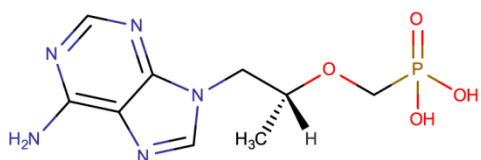
Tegaserod



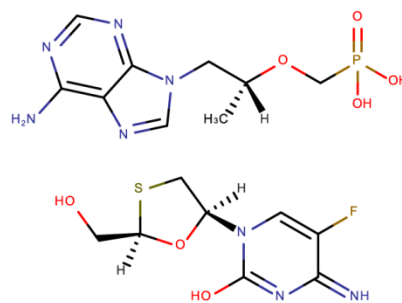
Telmisartan



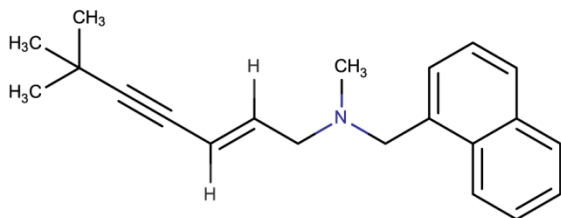
Temzolomide



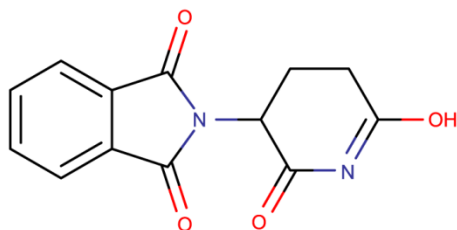
Tenofovir



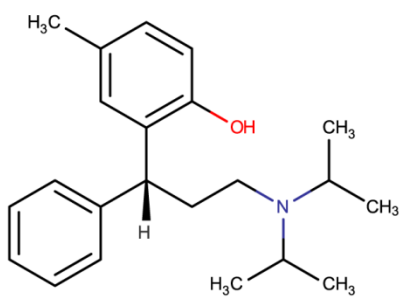
Tenofovir/Emtricitabine



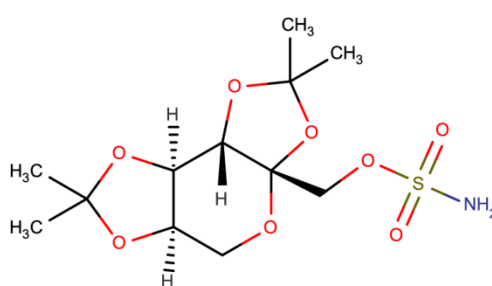
Terbinafine



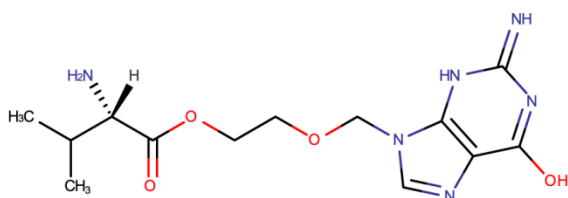
Thalidomide



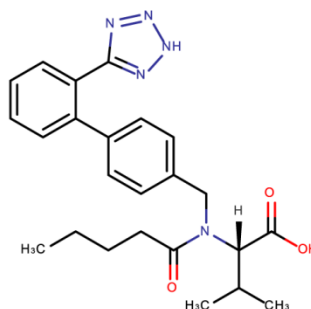
Tolterodine



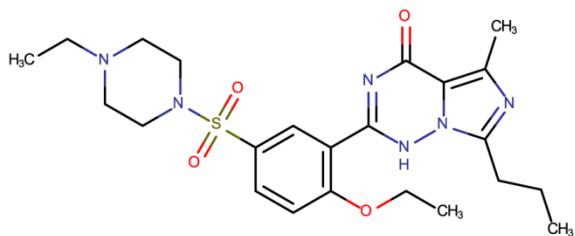
Topiramate



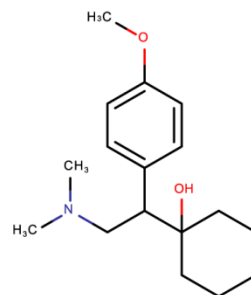
Valaciclovir



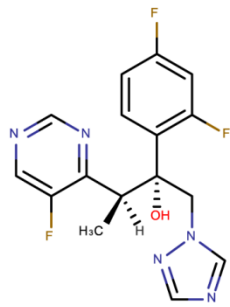
Valsartan



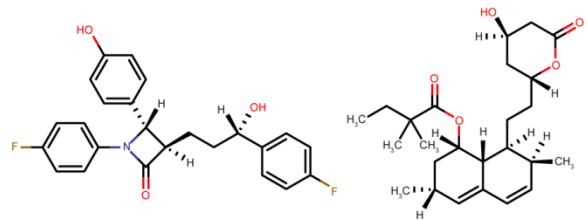
Vardenafil



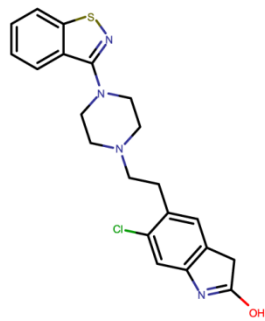
Venlafaxine



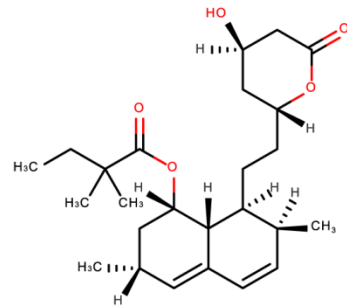
Voriconazole



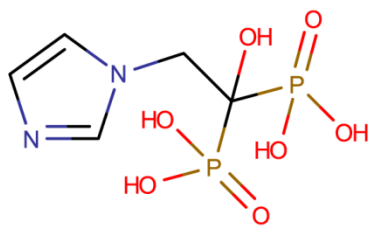
Vytorin



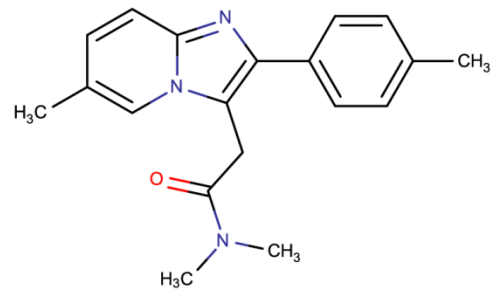
Ziprasidone



Zocor



Zolendronate



Zolpidem

### APPENDIX 3: SAMPLE SQL SOURCE FOR ChEMBL DATA EXTRACTION

#### Extract ChEMBL subset for comparison

```
CREATE TABLE PASS1 AS
select
  distinct
  a.TID as TID,
  a.ASSAY_ID as ASSAY_ID,
  b.ACTIVITY_ID as ACTIVITY_ID,
  b.molregno AS MOLREGNO,
  c.DOC_ID AS DOC_ID,
  b.STANDARD_TYPE AS STANDARD_TYPE,
  b.standard_value AS STANDARD_VALUE,
  b.standard_UNITS AS STANDARD_UNITS,
  b.ACTIVITY_COMMENT AS ACTIVITY_COMMENT

from
  chembl_14.ASSAY2TARGET a,
  chembl_14.ACTIVITIES b,
  chembl_14.docs c,
  chembl_14.ASSAYS d

where
  a.CONFIDENCE_SCORE > 2
  AND
  ( a.RELATIONSHIP_TYPE = 'D'
    or a.RELATIONSHIP_TYPE = 'H'
    or a.RELATIONSHIP_TYPE = 'M' )
  AND c.DOC_TYPE='PUBLICATION'
  AND d.ASSAY_TYPE='B'
  AND a.ASSAY_ID=b.ASSAY_ID
  AND a.ASSAY_ID=d.ASSAY_ID
  AND b.DOC_ID=c.doc_id
  AND
  ((UPPER(b.STANDARD_TYPE))='PKI'
   OR (UPPER(b.STANDARD_TYPE))= 'KI'
   OR (UPPER(b.STANDARD_TYPE))= 'LOG KI'
   OR (UPPER(b.STANDARD_TYPE))= '-LOG KI')
```

## Remove duplicated values

```
CREATE
  table pass2a as
  SELECT
    *
  FROM
    PASS1 a,
    (
      SELECT
        unique(miv)
      FROM
        (
          SELECT
            tid,
            molregno,
            min(activity_id) as miv ,
            standard_type ,
            standard_value ,
            standard_units
          FROM
            pass1
          WHERE
            standard_value IS NOT null
          GROUP BY
            tid,
            molregno,
            standard_type,
            standard_value,
            standard_units
          ORDER BY
            tid,
            molregno
        )
    ) b
  WHERE
    a.ACTIVITY_ID=b.miv
```

## Build final pairwise list

```
SELECT UNIQUE
  a.TID,
  targets.PREF_NAME,
  atjoin.COMPLEX,
  atjoin.MULTI,
  targets.ORGANISM,
  a.molregno,
  molecules.pref_name,
  a.doc_id,
  a.assay_id,
  a_assays.assay_type,
  a_assays.assay_organism,
  a.activity_id,
  a_activities.standard_type,
  a_activities.standard_value,
  a_activities.standard_units,
  a_activities.activity_comment,
  b.doc_id,
  b.assay_id,
  b_assays.assay_type,
  b_assays.assay_organism,
  b.activity_id,
  b_activities.standard_type,
  b_activities.standard_value,
  B_activities.standard_units

FROM
  PASS2 a,
  PASS2 b,
  CHEMBL_14.TARGET_DICTIONARY targets,
  CHEMBL_14.ASSAY2TARGET atjoin,
  CHEMBL_14.MOLECULE_DICTIONARY molecules,
  CHEMBL_14.DOCS a_docs,
  CHEMBL_14.ASSAYS a_assays,
  CHEMBL_14.ACTIVITIES a_activities,
  CHEMBL_14.DOCS b_docs,
  CHEMBL_14.ASSAYS b_assays,
  CHEMBL_14.ACTIVITIES b_activities

WHERE
  a.TID=b.TID
  AND a.MOLREGNO=b.MOLREGNO
  AND NOT(a.DOC_ID=b.DOC_ID)
  AND a.ACTIVITY_ID<b.ACTIVITY_ID
  AND a.TID=targets.TID
  AND a.TID=atjoin.TID
  AND a.ASSAY_ID=atjoin.ASSAY_ID
  AND a.MOLREGNO=molecules.MOLREGNO
  AND a.ACTIVITY_ID=a_activities.ACTIVITY_ID
  AND a.DOC_ID=a_docs.doc_id
  AND a.ASSAY_ID=a_assays.ASSAY_ID
```

AND a.DOC\_ID=a\_assays.DOC\_ID  
AND b.ACTIVITY\_ID=b\_activities.ACTIVITY\_ID  
AND b.DOC\_ID=b\_docs.doc\_id  
AND b.ASSAY\_ID=b\_assays.ASSAY\_ID  
AND b.DOC\_ID=b\_ASSAYS.DOC\_ID  
AND a.STANDARD\_VALUE IS NOT NULL  
AND b.standard\_value IS NOT NULL



## APPENDIX 4: SOURCE CODE FOR ATOM-PAIR DESCRIPTOR CALCULATION

```
#!/usr/bin/env python

import sys
import pybel
import networkx
import scipy
import csv

maxpath=15

atomlist=['C','N','O','P','S','F','Cl','Br','I','B']

atomlookup={6:'C', 7:'N', 8:'O', 15:'P', 16:'S', 9:'F', 17:'Cl', 35:'Br', 53:'I', 5:'B'}

offsets={'C':{'C':0, 'N':1, 'O':2, 'P':3, 'S':4, 'F':5, 'Cl':6, 'Br':7, 'I':8, 'B':9},\
         'N':{'C':1, 'N':10, 'O':11, 'P':12, 'S':13, 'F':14, 'Cl':15, 'Br':16, 'I':17, 'B':18},\
         'O':{'C':2, 'N':11, 'O':19, 'P':20, 'S':21, 'F':22, 'Cl':23, 'Br':24, 'I':25, 'B':26},\
         'P':{'C':3, 'N':12, 'O':20, 'P':27, 'S':28, 'F':29, 'Cl':30, 'Br':31, 'I':32, 'B':32},\
         'S':{'C':4, 'N':13, 'O':21, 'P':28, 'S':34, 'F':35, 'Cl':36, 'Br':37, 'I':38, 'B':39},\
         'F':{'C':5, 'N':14, 'O':22, 'P':29, 'S':35, 'F':40, 'Cl':41, 'Br':42, 'I':43, 'B':44},\
         'Cl':{'C':6, 'N':15, 'O':23, 'P':30, 'S':36, 'F':41, 'Cl':45, 'Br':46, 'I':47, 'B':48},\
         'Br':{'C':7, 'N':16, 'O':24, 'P':31, 'S':37, 'F':42, 'Cl':46, 'Br':49, 'I':50, 'B':51},\
         'I':{'C':8, 'N':17, 'O':25, 'P':32, 'S':38, 'F':43, 'Cl':47, 'Br':50, 'I':52, 'B':53},\
         'B':{'C':9, 'N':18, 'O':26, 'P':33, 'S':39, 'F':44, 'Cl':48, 'Br':51, 'I':53, 'B':54}}

pairscount=int((len(atomlist)*(len(atomlist)+1))/2)

def mol_to_networkxgraph(mol):
    edges = []
    bondorders = []
    for bond in pybel.ob.OBMolBondIter(mol.OBMol):
        bondorders.append(bond.GetBO())
        edges.append( (bond.GetBeginAtomIdx() - 1, bond.GetEndAtomIdx() - 1) )
    g = networkx.Graph()
    g.add_edges_from(edges)
    return g

def check_atoms(mol):
    doespass = True
    for atm in molecule.atoms:
        if atm.atomicnum not in atomlookup.keys():
            doespass=False
            break
    return(doespass)

fileheader=[]
fileheader.append('Molecule')
for l in range(maxpath):
    for i in range(0,len(atomlist)):
        for j in range(i,len(atomlist)):
            label=atomlist[i]+repr(l+1)+atomlist[j]
```

```

fileheader.append(label)

if len(sys.argv)< 3:
    print "Usage: apdescs.py infile.sdf outfile.csv"
    exit(64)

outfile=open(sys.argv[2],'wb')
outhand=csv.writer(outfile,quoting=csv.QUOTE_NONNUMERIC)

outhand.writerow(fileheader)

molkount=1

for molecule in pybel.readfile("sdf",sys.argv[1]):
    molecule.removeh()
    msize=len(molecule.atoms)

    goodmol=check_atoms(molecule)
    if goodmol==False:
        print "Molecule",repr(molkount),'has an invalid atom type. Skipping.'
        break

    if molecule.title:
        molname=molecule.title.replace(' ',';')
    else:
        molname="Molecule "+repr(molkount)

    kountarray=scipy.zeros((pairskount*maxpath))

    graph=mol_to_networkxgraph(molecule)

    dists=scipy.zeros((msize,msize))

    for i in range(msize):
        for j in range(i+1,msize):
            dists[i,j]=len(networkx.shortest_path(graph,i,j))-1
            dists[j,i]=dists[i,j]

    for i in range(msize):
        for j in range(i+1,msize):
            atoma=atomlookup[molecule.atoms[i].atomicnum]
            atomb=atomlookup[molecule.atoms[j].atomicnum]
            shortestdist=int(dists[i,j])
            if shortestdist>maxpath:
                print "In molecule",molkount,"distance between atoms",atoma,"and",atomb,"is
over",maxpath,"bonds. Skipping"
                continue
            discoffset=((shortestdist-1)*pairskount)+offsets[atoma][atomb]
            kountarray[discoffset]=kountarray[discoffset]+1

    outputlist=[molname]
    for element in kountarray:
        outputlist.append(int(element))

```

```
outhand.writerow(outputlist)
```

```
molcount=molcount+1
```

## APPENDIX 5: SOURCE CODE FOR VARIABLE SELECTION QSAR MODEL CONSTRUCTION

```
library(caret)

###
### This script is configured currently to generate classifier models using rSVM
###
### In order to switch to another classifier method:
### substitute another method into the buildModel function below
### if you are using RF models, you may want to modify the modelCtrl command
### to use OOB instead of 5-fold cross validation for parameter optimization
###
### In order to switch to a regression modeling scheme, more changes are needed:
### another method must be substituted in the buildModel function
### the modelCtrl command should be switched to use R-squared for internal opt.
### Rsquared should be substituted in for accuracy in all train and all
### GetTrainPerf function calls
### Calls to makeCCR should be replaced with calls to makeR2
### Any changes made to the main driving routine that reference additional values
### returned by the makeCCR function will need to be changed (R2 and Ro2 instead
### of CCR, sensitivity and specificity)
###

options(warn=1)

### uncomment the next line if reproducible runs are needed for testing
#set.seed(301)

### Major Control Variables are Set Below
###
### modelKount how many distinct candidates to consider at once
### modelSize how many descriptors in each candidate model
### generations how many cycles of GA optimization to run
### respVar which column of data has the activity class or pKi
### firstData which is the first column containing descriptor values
### extFolds how many folds to use in external cross-validation
### intCutoff what is the numerical threshold for accepting a model
### after internal cross validation
### extCutoff what is the numerical threshold for accepting a model
### after external cross-validation
### returnedModels how many models are returned after descriptor selection
### for each fold
### mutateProb reciprocal of the probability that a single candidate
### model will be mutated in any given cycle
###

modelKount<-8
modelSize<-15
generations<-100
respVar<-2
```

```

firstData<-5
extFolds<-5
intCutoff<-0.7
extCutoff<-0.6
returnedModels<-2
maxCorr<-0.9
mutateProb<-10

### Utility Functions

initModel <- function(gk,range) {
  return( sort(sample(1:range,size=gk)))
}

buildModel <- function(dataIn) {
  trialFit<-train(trainingResps ~
  .,data=dataIn,method="svmRadial",metric="Accuracy",maximize=TRUE,tunelength=9,trControl=modelCtrl)
  return(trialFit)
}

checkModel <- function(foo) {
  newchrome<-foo
  while (sum(duplicated(newchrome)==TRUE)!=0) {
    newchrome[which(duplicated(newchrome))[1]]<-sample(1:modelSize,1)
    newchrome<-sort(newchrome)
  }
  if (!(identical(newchrome,foo))) {
    retval<-newchrome
  } else {
    retval<-TRUE
  }
  return(retval)
}

makeCCR <-function(extActual,extPreds) {
  contTable<-table(extActual,extPreds)
  sens<-contTable[1,1]/(contTable[1,1]+contTable[1,2])
  spec<-contTable[2,2]/(contTable[2,1]+contTable[2,2])
  CCR<-(sens+spec)/2
  retVal<-c(CCR,sens,spec)
  return(retVal)
}

plotPerf <-function(trace) {
  plot(trace[2,],col="black",ylim=c(0,0.95),type="l")
  lines(trace[3,],col="orange",type="l")
  lines(trace[4,],col="green",type="l")
}

modelCtrl<-trainControl(method='repeatedCV',number=5,repats=5)

```

```

### Read Data

rawData<-read.table("FullAlpha2aDescriptors.csv",header=TRUE,row.names=1,sep=',')

### Preclear Data and Generate Folds
### any other filtering or subsetting should occur here

badResps<-which(is.na(rawData[,1]))
rawData<-rawData[-badResps,]
dataFolds<-createFolds(rawData[,1],k=extFolds)

rawDescs<-rawData[,firstData:(length(rawData[,1]))]
rawResps<-rawData[,respVar]
rawNames<-names(rawDescs)

### Build an Optimized Subset Model

buildOptModel <- function(extFold) {

  rawExtDescs<-rawDescs[dataFolds[[extFold]],]
  externalResps<-rawResps[dataFolds[[extFold]]]
  rawTrainDescs<-rawDescs[-dataFolds[[extFold]],]
  trainingResps<-rawResps[-dataFolds[[extFold]]]
  foldNames<-rawNames

  nZV<-nearZeroVar(rawTrainDescs,uniqueCut=5)
  rawTrainDescs<-rawTrainDescs[,-nZV]
  rawExtDescs<-rawExtDescs[,-nZV]
  foldNames<-foldNames[,-nZV]

  corMat<-cor(rawTrainDescs)
  highCorr<-findCorrelation(corMat,cutoff=maxCorr)
  rawTrainDescs<-rawTrainDescs[,-highCorr]
  rawExtDescs<-rawExtDescs[,-highCorr]
  foldNames<-foldNames[,-highCorr]
  nDesc<-length(foldNames)

  scaleFactors<-preProcess(rawTrainDescs,method=c('center','scale'))
  trainingDescs<-predict(scaleFactors,rawTrainDescs)
  externalDescs<-predict(scaleFactors,rawExtDescs)

  modelDescs<-as.vector(NULL)
  modelScores<-as.vector(NULL)
  modelStore<-as.list(NULL)

  for (initKount in seq(modelKount)) {
    modelStore<-append(modelStore,o)
    newModel<-initModel(modelSize,nDesc)
    modelDescs<-rbind(modelDescs,newModel)
    trainingInput<-cbind(trainingResps,trainingDescs[modelDescs[initKount,]])
    initFit<-buildModel(trainingInput)
  }
}

```

```

modelScores<-c(modelScores,getTrainPerf(initFit)$TrainAccuracy)
modelStore[[initKount]]<-initFit
}

### Check for duplication between models
while ( sum(duplicated(modelDescs)!=0) ) {
  dupedModel<-which(duplicated(modelDescs)==TRUE)[1]
  modelDescs[dupedModel,]<- initModel(modelSize,nDesc)
  trainingInput<-cbind(trainingResps,trainingDescs[modelDescs[dupedModel,]])
  newFit<-buildModel(trainingInput)
  modelScores[dupedModel]<-getTrainPerf(newFit)$TrainAccuracy
  modelStore[[dupedModel]]<-newFit
}

localTrace<-NULL
for (iters in seq(generations)) {

  ### Check for duplicate descriptors in each model
  for (j in seq(modelKount)) {
    unpack<-checkModel(modelDescs[j,])
    if (length(unpack)==modelSize) {
      modelDescs[j,]<-unpack
      trainingInput<-cbind(trainingResps,trainingDescs[,modelDescs[j,]])
      newFit<-buildModel(trainingInput)
      modelScores[j]<-getTrainPerf(newFit)$TrainAccuracy
      modelStore[[j]]<-newFit

    }
  }

  if (iters%%25==0) {
    cat(c('Iteration ',iters[1],'\n'))
  }

  parents<-sort(sample(modelKount,2))

  bpnt<-sample((1:(modelSize-1)),1)

  c1<-sort(c(modelDescs[parents[1],((1):(bpnt))],modelDescs[parents[2],((bpnt+1):(modelSize))]))
  c2<-sort(c(modelDescs[parents[2],((1):(bpnt))],modelDescs[parents[1],((bpnt+1):(modelSize))]))

  if (sample(1:mutateProb,1)==2) {
    newc1<-c1
    newc1[sample(modelSize,1)]<-sample(1:nDesc,1)
    c1<-sort(newc1)
  }

  if (sample(1:mutateProb,1)==3) {
    newc2<-c2
    newc2[sample(modelSize,1)]<-sample(1:nDesc,1)
    c2<-sort(newc2)
  }
}

```

```

}

unpack<-checkModel(c1)
if (length(unpack)==modelSize) {
  c1<-unpack
}
trainingInput<-cbind(trainingResps,trainingDescs[,c1])
c1model<-buildModel(trainingInput)
c1score<-getTrainPerf(c1model)$TrainAccuracy

unpack<-checkModel(c2)
if (length(unpack)==modelSize) {
  c2<-unpack
}
trainingInput<-cbind(trainingResps,trainingDescs[,c2])
c2model<-buildModel(trainingInput)
c2score<-getTrainPerf(c2model)$TrainAccuracy

kids<-rbind(c1,c2)
kidscores<-c(c1score,c2score)
kidmodels<-as.list(1:2)
kidmodels[[1]]<-c1model
kidmodels[[2]]<-c2model

### figure out rank order for models
if (c1score != c2score) {
  if ( kidscores[1] < kidscores[2] ) {
    cmax<-2
    cmin<-1
  } else {
    cmax<-1
    cmin<-2
  }
} else {
  cmin<-1
  cmax<-2
}

if (modelScores[parents[1]] != modelScores[parents[2]]) {
  if (modelScores[parents[1]]<modelScores[parents[2]]) {
    pmin<-parents[1]
    pmax<-parents[2]
  } else {
    pmin<-parents[2]
    pmax<-parents[1] }
} else {
  pmin<-parents[1]
  pmax<-parents[2]
}
}

```



```

### Replace Parents with Children if appropriate
if ( kidscores[cmax] >= modelScores[pmax] ){
  if ( kidscores[cmin] >= modelScores[pmin] ) {
    modelDescs[pmax,]<-kids[cmax,]
    modelDescs[pmin,]<-kids[cmin,]
    modelScores[pmax]<-kidscores[cmax]
    modelScores[pmin]<-kidscores[cmin]
    modelStore[[pmax]]<-kidmodels[[cmax]]
    modelStore[[pmin]]<-kidmodels[[cmin]]
  } else {
    modelDescs[pmin,]<-kids[cmax,]
    modelScores[pmin]<-kidscores[cmax]
    modelStore[[pmin]]<-kidmodels[[cmax]]
  }
} else {
  if (kidscores[cmax] >= modelScores[pmin]) {
    modelDescs[pmin,]<-kids[cmax,]
    modelScores[pmin]<-kidscores[cmax]
    modelStore[[pmin]]<-kidmodels[[cmax]]
  }
}

### Check for duplication in candidate models
while (sum(duplicated(modelDescs))==TRUE) {
  dupe<-(which(duplicated(modelDescs))==TRUE)[1]
  modelDescs[dupe,]<- initModel(modelSize,nDesc)
  trainingInput<-cbind(trainingResps,trainingDescs[,modelDescs[dupe,]])
  modelStore[[dupe]]<-buildModel(trainingInput)
  modelScores[dupe]<-getTrainPerf(modelStore[[dupe]])$TrainAccuracy

}

### Keep track of optimization performance for this cycle
localTrace <- cbind(localTrace,c(iters,min(modelScores),mean(modelScores),max(modelScores)))
}

# trainingInput<-cbind(trainingResps,trainingDescs)
# trialModel<-train(trainingResps ~ .,
data=trainingInput,method="svmRadial",metric="Accuracy",maximize=TRUE,tuneLength=5,trControl=modelCtrl)

### Construct a return value containing all the needed information
###
### Elements in the returned list are:
###
### 1 Training descriptors used for this fold
### 2 Descriptor names used in this fold
### 3 The normalization function for this fold
### 4 Names of molecules in the external set
### 5 Optimization performance for this fold
### 6 List of actual model objects from train call
### 7 List of internal validation values

```

```

### 8 List of external validation values
### 9 List of external validation set prediction vectors
###

retval<-as.list(sequence(9))
retmods<-as.list(seq(returnedModels))
retint<-seq(returnedModels)
retext<-seq(returnedModels)
retExp<-as.list(seq(returnedModels))
modelOrder<-order(modelScores,decreasing=TRUE)
retval[[1]]<-trainingDescs
retval[[2]]<-foldNames
retval[[3]]<-scaleFactors
retval[[4]]<-row.names(externalDescs)
retval[[5]]<-localTrace
for (modno in seq(returnedModels)) {
  retmods[[modno]]<-modelStore[[modelOrder[modno]]]
  retint[[modno]]<-getTrainPerf(retmods[[modno]])$TrainAccuracy
  testValues<-predict(retmods[[modno]],externalDescs)
  modelEval<-makeCCR(externalResps,testValues)
  retext[[modno]]<-modelEval[1]
  retExp[[modno]]<-seq(length(testValues))
  retExp[[modno]]<-testValues
}
retval[[6]]<-retmods
retval[[7]]<-retint
retval[[8]]<-retext
retval[[9]]<-retExp
return(retval)
}

```

### Main Driving Loop

```

rawOutput<-as.list(seq(extFolds))

# Maybe try to use foreach to parallelize?

for (efold in seq(extFolds)) {
  cat(c('Building fold ',efold,'\n'))
  rawOutput[[efold]]<-buildOptModel(efold)
}

```