

ADVANCING QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP
STRATEGIES IN LIGAND-BASED COMPUTER-AIDED DRUG DESIGN

By

Mariusz Butkiewicz

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

In

Chemistry

August, 2014

Nashville, Tennessee

Approved:

Jens Meiler, Ph.D.

Brian O. Bachmann, Ph.D.

David W. Wright, Ph.D.

Clare M. McCabe, Ph.D.

Copyright © 2014 by Mariusz Butkiewicz

All Rights Reserved

DEDICATION

To my parents, my sister, and Nicole.

ACKNOWLEDGEMENTS

Over the past years, I have received support and encouragement from a great number of individuals to whom I am very grateful.

I would like to express my deepest and sincere gratitude to my advisor, Dr. Jens Meiler. Coming to Nashville and joining the Meiler laboratory to start my graduate studies has been a tremendous opportunity and extraordinary experience in my life. Jens was an excellent mentor and supported me on each step in my graduate career. His guidance taught me how to approach scientific problems, how to ask-the right scientific questions, and how to write and present scientific work. Jens found the right balance between encouraging my own scientific explorations and providing invaluable guidance and help. I would like to thank Dr. Meiler for making the past several years such a pleasant academic experience.

The members of my dissertation committee, Dr. David Wright, Dr. Brian Bachmann, and Dr. Clare McCabe, were a great source of support and guidance for my graduate work. Their insightful comments and constructive criticism gave appreciated impulses to my research.

Many friends and colleagues in the Meiler lab provided great help and support during the past years. Particularly, I would like to thank Will Lowe, Jeff Mendenhall, Nils Woetzel, Ralf Mueller, and Kristian Kaufmann for great discussions, ideas, and inspiration.

I would like to express my deepest gratitude to my family, my parents, and my sister, Sylvia. Their constant love and support made this incredible experience possible, despite the large geographical distance.

Finally, I want to thank my better half, Nicole Restrepo. Her unconditional love and support provided balance in my life and made this dissertation possible. She made sure that I took care of myself during my years in graduate school.

TABLE OF CONTENTS

	Page
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	viii
LIST OF ABBREVIATIONS	x
SUMMARY	xii
Chapter	
1. INTRODUCTION	1
Distinct CADD strategies cope with availability of protein structure information.....	1
Quantitative Structure Activity Relationships correlate chemical structure to biological activity	3
Molecular descriptors encode chemical structure in QSAR modeling	3
Machine learning approaches model QSARs between chemical structure and biological activity	4
Improvements to prediction accuracy of QSAR models.....	5
High-throughput screening sets knowledge base for virtual screening in drug discovery	6
Curating and identifying publicly and commercially available biological assay data sets	6
Virtual high-throughput screening expands chemical search space in comparison to traditional HTS....	7
Allosteric potentiators of mGlu ₅ provide a novel approach for treatment of schizophrenia.....	7
Novel inhibitors for <i>Plasmodium falciparum</i> have potential to diminish Malaria drug resistance.....	8
Predicting molecular properties through Quantitative Structure Property Relationships	9
Concepts of CADD are implemented in Bio Chemistry Library	10

2. LIGAND-BASED VIRTUAL HIGH-THROUGHPUT SCREENING BENCHMARK AND METHOD DEVELOPMENT	11
Introduction.....	11
Results and Discussion.....	15
Experimental.....	22
Conclusions.....	34
3. IDENTIFICATION OF PATHWAY SPECIFIC INHIBITORS FOR β -HEMATIN CRYSTALLIZATION IN PLASMODIUM FALCIPARUM INVOLVED IN MALARIA.....	36
Introduction.....	36
Results.....	38
Methods.....	44
Discussion.....	47
4. NOVEL ALLOSTERIC MODULATORS FOR MGLU ₅ RELATED TO CPPHA BINDING	50
Introduction.....	50
Results.....	54
Methods.....	61
Discussion.....	65
Conclusions.....	67
5. SMALL MOLECULE PROPERTY PREDICTIONS	69
Comparative Analysis of Machine Learning Techniques for the Prediction of LogP	69
Quantitative Structure Property Modeling for the Prediction of DMPK Parameters Intrinsic Clearance and Plasma Protein Binding.....	80
6. DISCUSSION	90

Conclusions and future directions.....	90
Methods development and benchmarking of Quantitative Structure Activity Relationships	90
Discovery of pathway specific antimalarial hit compounds to diminish <i>P. falciparum</i>	92
Selective allosteric modulators for distinct mGlu ₅ site related to CPPHA binding.....	93
Molecular property predictions	94
Future perspectives of LB-CADD	94
APPENDIX.....	100
General Comments.....	100
Supporting information for Chapter 2.....	100
Supporting information for Chapter 3.....	113
Supporting information for Chapter 4.....	115
Supporting information for Chapter 5.....	117
Tutorial for QSAR modeling, virtual screening, and cluster analysis with BCL::ChemInfo	119
Implementation of multi-output Support Vector Regression.....	132
Implementation of an auto-encoder algorithm for molecular descriptor compression	134
Mapping of protein-ligand interactions elucidates determinants of structure-activity relations	137
Multi-output prediction QSAR model to create a small molecule prediction profile for mGlu _{1,8}	140
REFERENCES	141

LIST OF TABLES

Table 1: Overview of PubChem biological assays and data set statistics.....	16
Table 2: Overview of descriptor selection results.....	19
Table 3: Overview of consensus benchmark results for all PubChem datasets.....	21
Table 4: Listing with all available datasets used for QSAR modeling and virtual screening study.....	54
Table 5: Seven confirmed mGlu5 modulators predicted to interact with CPPHA binding related site.....	60
Table 6: The original molecular descriptors by category.....	76
Table 7: Model statistics for best predictors.....	78
Table 8: Consensus Predictors.....	79
Table 9: Data set composition.....	85
Table 10: Molecular descriptors by category.....	86
Table 11: Optimized parameters.....	87
Table 12: Model correlation results ($r_{p/s}/\text{rmsd}$) for independent validation set.....	87
Table 13: Overview of descriptors by category, description, and number of descriptor features.....	101

LIST OF FIGURES

Figure 1: Overview of computer aided drug design / discovery.....	2
Figure 2: Dendrogram representing the majority of cluster scaffolds of 530 β -hematin inhibitors.....	39
Figure 3: QSAR model predictions compared to naïve substructure search.....	43
Figure 4: Validation of β -hematin inhibitors in Plasmodium falciparum.....	47
Figure 5: Clustering of compounds 130 actives and 145 inactives in the CPPHA series.....	55
Figure 6: Schematics of data set composition for each of the three QSAR models hit compounds.....	56
Figure 7: Listing of ROC curve quality measures for each cross-validated model on a log scale.....	57
Figure 8: Distribution plot shows the similarity of ordered compounds compared to known PAMs.....	59
Figure 9: Experimental validation of the three most active compounds by calcium binding.....	61
Figure 10: Confirmation of CPPHA site specific binding through radio ligand binding assay.....	62

Figure 11: Schematic view of an ANN.....	71
Figure 12: Schematic view of a Support Vector ϵ – tube.	72
Figure 13: Schematic view of kNN cluster centers with its determined nearest neighbor environments...	74
Figure 14: Correlation plot of best consensus predictions compared to experimental data.....	78
Figure 15: ANN, SVM, kNN, and XlogP predictions compared to Reaxys/MDDR experimental values.	79
Figure 16: Overview of correlation plots for respective consensus models	88
Figure 17: Pair-wise comparison of optimized descriptor sets.	103
Figure 18: Heat maps reporting Tanimoto coefficients of descriptor overlap.	104
Figure 19: The TNR-TPR curve is shown for SAIDs 488997, 485290, and 2258.	106
Figure 20: ROC curve evaluating the consensus model.	111
Figure 21: The heme speciation assay conducted on 17 β -hematin inhibiting GSK compounds.	112
Figure 22: Graphical overview of molecular descriptors applied in the tutorial.	123
Figure 23: Explanatory schematic of variables used in multi-output SVR training.	134
Figure 24: Schematic view of training de-noising autoencoders.	137
Figure 25: Schematic view of protein-ligand interaction.....	139

LIST OF ABBREVIATIONS

QSAR	-	Quantitative Structure Activity Relations/Relationships
QSPR	-	Quantitative Structure Property Relations/Relationships
BCL	-	BioChemistryLibrary
HTS	-	high-throughput screening
vHTS	-	virtual high-throughput screening
CADD	-	computer-aided drug design
SB-CADD	-	structure-based computer-aided drug design
LB-CADD	-	ligand-based computer-aided drug design
logP	-	water/octanol partition coefficient
ANN	-	artificial neural network
SVM	-	support vector machine
kNN	-	kappa nearest neighbor algorithm
KN	-	kohonen network
DT	-	decision tree
mGlu	-	metabotropic glutamate receptor
mGlu5	-	metabotropic glutamate receptor subtype 5
GPCR	-	G protein coupled receptor
GPU	-	graphics processing unit
AID	-	PubChem bioassay identifier
SAID	-	PubChem summary bioassay identifier
DFB	-	1-(3-fluorophenyl)-N-((3-fluorophenyl)-methylideneamino)-methanimine
CDPPB	-	3-cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl)-benzamide
CPPHA	-	N-(4-chloro-2-((1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)methyl)phenyl)-2-hydroxybenzamide
SAR	-	structure–activity relationship

MPEP	-	2-methyl-6-(phenylethynyl)-pyridine
VU-29	-	4-nitro-N-(1,3-diphenyl-1H-pyrazol-5-yl)-benzamide
5MPEP	-	5-methyl-2-(phenylethynyl)-pyridine
NCFP	-	N-(4-chloro-2-((4-fluoro-1,3-dioxoisindolin-2-yl)methyl)phenyl)picolinamide
NAM	-	negative allosteric modulator
PAM	-	positive allosteric modulator
ROC	-	receiver operating characteristics curve
ENR	-	quality measure enrichment
PPV	-	positive predictive value
FPR	-	false positive rate
DMPK	-	drug metabolism and pharmacokinetics
CL_{int}	-	intrinsic clearance
f_u	-	plasma protein binding as the fraction of unbound compound

SUMMARY

The focus of this thesis was to develop a virtual screening framework that allows for small molecule data handling, QSAR model development, optimization and analysis, as well as subsequent QSAR model prediction through virtual screening as an integral part of collaborative drug discovery projects. All algorithms are implemented in the software suite BCL::ChemInfo which is part of the BioChemistry library (BCL), an in-house object-oriented library providing functionality to model small molecules and proteins.

Chapter I provides an introduction to the field of computer-aided drug design/discovery, describes the underlying concepts of BCL::ChemInfo, and highlights two biological systems where the here introduced software framework contributed significantly to respective drug discovery efforts.

Chapter II details an overview and benchmark of the BCL::ChemInfo method development. This chapter is largely a reproduction of the publication Mariusz Butkiewicz, Edward W. Lowe, Ralf Mueller, Jeffrey L. Mendenhall, Pedro L. Teixeira, C. David Weaver, and Jens Meiler titled "Benchmarking ligand-based virtual high-throughput screening with the PubChem database." *Molecules* 18, no. 1 (2013): 735-756. This work represents the core of this thesis and is crucial to all application projects involving BCL::ChemInfo.

Chapter III describes the application of BCL::ChemInfo towards finding antimalarial inhibitors regarding a specific pathway of action. It articulates a collaborative effort between the laboratories of Dr. David Wright and Dr. Jens Meiler at Vanderbilt to guide experimental expertise with computational algorithms to find effective therapeutics against Malaria. The underlying manuscript is in the stage of finalization and will be submitted shortly. It is co-authored by Rebecca D. Sandlin and Mariusz Butkiewicz. Mariusz Butkiewicz was guiding the computational component of the project involving the QSAR modeling, virtual screening, and *in silico* analysis. Rebecca D. Sandlin provided the initial HTS data for identified β -

hematin inhibitors and validated the computational findings experimentally. The respective sections provided exclusively by Rebecca D. Sandlin are marked in this chapter.

Chapter IV summarizes the discovery of selective allosteric modulators of mGlu₅ for a novel specific binding site applying BCL::ChemInfo. It is a collaborative research project between the laboratories of Dr. P. Jeffrey Conn, Dr. Craig W. Lindsley, and Dr. Jens Meiler at Vanderbilt. The resulting manuscript highlights computational experiments that are explained and discussed in this chapter. Mariusz Butkiewicz performed the computational work and is first author of the manuscript which is expected to be submitted, shortly. Experimental validations of the research finding were performed by Alice L. Rodriguez. Respective sections provided exclusively by Alice L. Rodriguez are marked in this chapter.

Chapter V investigates an alternative field of application for BCL::Cheminfo involving the prediction of molecular properties in drug discovery. This chapter is based on the publications Edward W. Lowe, Mariusz Butkiewicz, Matthew Spellings, Albert Omlor, and Jens Meiler, "Comparative analysis of machine learning techniques for the prediction of logP.", *IEEE Symposium in Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2011, pp. 1-6. and Edward W. Lowe, Mariusz Butkiewicz, Zollie White III, Matthew Spellings, Albert Omlor, Jens Meiler, "Comparative Analysis of Machine Learning Techniques for the Prediction of the DMPK Parameters Intrinsic Clearance and Plasma Protein Binding", *Proceedings of 4th International Conference on Bioinformatics and Computational Biology 2012*, p. 25, 2012. As second author to both citations, Mariusz Butkiewicz contributed to the writing aspect of both publications as well as performing a substantial part of the involved computational experiments.

Finally, chapter VI provides a discussion about the previous chapters and summarizes individual perspectives of research. An outlook about future work is presented showing possible next steps to guide improvement of BCL::ChemInfo. Finally, the appendix gives detailed information about side projects and ideas that could not be discussed in the chapters and have not been published. More detailed information

about command lines, scripts, data, and models provided in the appendix can be found on a DVD that accompanies this thesis.

1. INTRODUCTION

The investigation of signal transduction processes through receptor proteins and the understanding of their interaction with small molecule ligands are key elements of fundamental biomedical research and drug discovery. Historically, active ingredients were discovered through traditional therapies or more serendipitous findings. With the evolution of technology, libraries of small organic molecules, natural products, and extracts are evaluated in an automated and systematic fashion. High-throughput screening (HTS) was introduced to find novel compounds by rapidly executing assay experiments against single biological targets associated with disease altering phenotypes [4]. At the same time, interest in computer-aided drug discovery/design (CADD), a purely computational technology, increased significantly [5]. Traditional HTS experiments often result in multiple hit compounds which some can be further optimized to lead compounds and probe molecules. The typical hit rate is very low limiting the applicability of HTS in research efforts attempting evaluating large compound libraries [6]. CADD provides an alternative solution to this problem. Large compound libraries that are not readily accessible to traditional HTS can be screened *in silico*. This substantially reduces the number of compounds necessary to experimentally screen, while the same level of lead compound discovery is retained. Thus, the chemical space is extended to identify possible hit compounds. Molecules predicted to be active can be prioritized while molecules predicted to be inactive can be discarded. It significantly decreases the cost and time commitment of a full HTS experiment while maintaining a comparable lead discovery rate. Furthermore, CADD requires substantially less preparation time. As a technology to guide HTS studies, CADD campaigns can be applied to virtually screen compounds while the traditional HTS assay is being prepared.

Distinct CADD strategies cope with availability of protein structure information

In CADD, two strategies (Structure-based, computer-aided drug design (SB-CADD) and ligand-based computer-aided drug design/discovery (LB-CADD)) are available for drug discovery depending on the availability of protein target structural data. An overview is given in Figure 1. SB-CADD methods depend

on the availability of a structural model determined by X-ray crystallography or NMR techniques. If the target structure is known, SB-CADD can be applied through approaches such as ligand docking, de novo design, molecular dynamics, or pharmacophore modeling [7]. However, a large fraction of proteins targeted by therapeutics have no suitable structural model for structure-based virtual screening. Alternatively, LB-CADD can be applied when only ligand structure information and its associated biological response with a protein target are given. One central paradigm leverages properties and determinants of chemical structures and the associated biological response to describe its relation. Involved techniques include quantitative structure activity relationships (QSAR), virtual screening, and pharmacophore modeling. Pharmacophore modeling can be categorized in both sub fields. In SB-CADD, pharmacophore models take steric and electronic features of the binding site into account that are important for ligand binding. In LB-CADD, a QSAR model can be applied to determine a pharmacophore map of the ligand independent of the protein target structure.

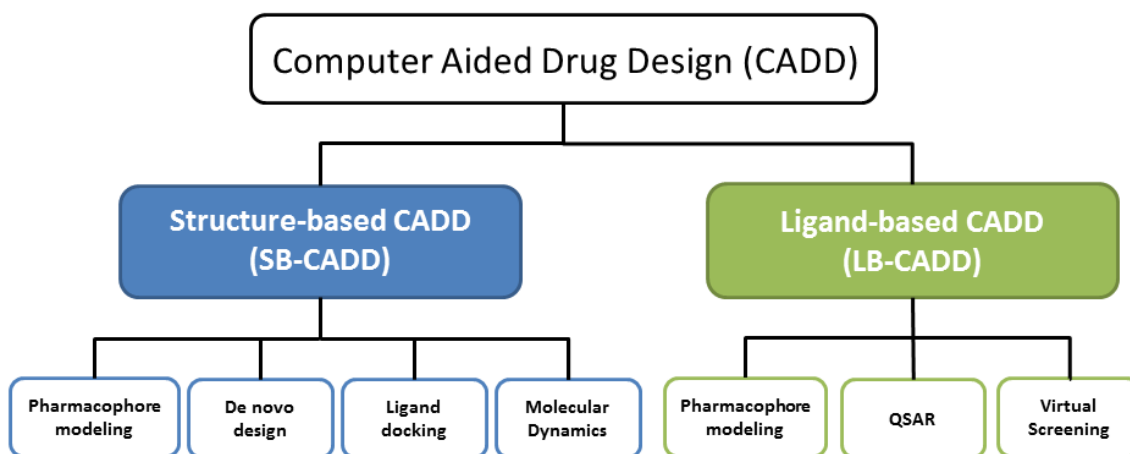


Figure 1: Overview of computer aided drug design / discovery.

CADD is divided into two main categories. Depending on the availability of determined structural information of a protein two different approaches can be chosen: SB-CADD and LB-CADD. Involved techniques are listed for each category.

Quantitative structure activity relationships correlate chemical structure to biological activity

The ligand-based approach employs quantitative structure–activity relationships (QSAR), a technology that does not require information about the biological target structure. QSAR is an area of computational research that builds virtual models to predict the biological activity, the binding affinity, or the toxic potential of existing or hypothetical molecules [8]. They seek to quantitatively correlate complex non-linear relations between chemical and physical properties of a molecule with its biological response, such as activity for a specific biological target using statistical models. The application of QSARs plays an important role in a variety of fields such as drug discovery [9] and toxicity predictions [10]. Knowledge acquired from compounds already biologically tested direct the design of future molecules.

Molecular descriptors encode chemical structure in QSAR modeling

Molecular descriptors play a fundamental role in encoding chemical information contained in a molecule [11]. Through mathematical procedures the chemical information contained within a symbolic representation of a molecule is transferred into a numerical code. Two main categories arise containing experimental measurements, such as the octanol/water partition coefficient ($\log P$), molar refractivity, dipole moment, polarizability, and secondly, theoretical molecular descriptors derived from such symbolic representations of molecules. In small focused chemical libraries theoretical descriptors are significant in ‘fragment-independent’ encoding schemes. Examples include autocorrelation functions which encode identity and electronic attributes of molecular structure (i.e. atom types, partial atomic charges, electronegativity and polarizability) into vector representations. Radial distribution functions describe how the atomic density varies as a function of the distance from one particular atom. Surface area correlation functions store molecular shape geometry for molecules with known biological activity into machine learning models. Each encoding function is independent of translation or rotation of the molecule. They can be readily applied to conformational ensembles and yield a numerical description of constant length independent from size and composition of the substance. A third-party 3D conformation

generator, CORINA [12], allows for 3D coordinates generation for molecules prior to descriptor calculation.

Machine learning approaches model QSARs between chemical structure and biological activity

Machine learning approaches have shown exciting potential in estimating biological target data. In recent years the potential of approaches such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) for establishing highly non-linear relations has become apparent [13-17]. The algorithms learn to recognize complex patterns and make intelligent decisions based on an established compound library. Imposing these sets of patterns obtained from the learning process, the algorithms are able to recognize not yet tested molecules and categorize them towards a given outcome. Many different technologies can be applied in the QSAR domain including SVMs, ANNs, Kohonen networks (KN), kappa nearest neighbor algorithm (kNN), and Decision Trees (DT). SVMs [18-21] use distinct mathematical functions to transform a given set of molecular descriptors of the chemical structure into a hyperspace optimized to separate a set of data points according to a given, usually binary property. This is achieved by introducing a hyperplane in this new hyperspace. Modified versions are also able to handle regression problems like function approximation. ANNs consist of an interconnected group of artificial neurons and process information adaptively using restricted set of transformations. ANNs have been used for several years in chemistry and biochemistry to describe QSAR [22]. Structurally related to ANNs are KNs, also known as self-organizing maps. KNs are trained in an unsupervised fashion and pursue the discretization of the training input into a network or map [23-25]. A simpler unsupervised learning algorithm is represented by kNNs. Based on the principle of majority voting, kNNs resample a type of clustering through a given distance metric defining the neighborhood of a given feature vector, and thus guide the prediction outcome [26, 27]. DTs represent a sequence of branching operations that break up complex problems into a union of simplified sub-problems. Decision analysis is based on knowledge of statistical decision theory. Key differences of SVMs, ANNs and DTs are an increased reproducibility for SVMs, a slightly higher flexibility for ANNs [28] and a more comprehensible model representation for

DTs, kNNs, and KNs. Their critical advantage compared to linear methods such as multiple linear regression lies in the flexibility of the models. ANN and SVM models can adapt to complex interrelations and are capable of detecting small signals at high noise levels. Hence, each of the methods is applied when no simple mathematical model can be assumed, many potential parameters interact, and the experimental uncertainties are high.

Improvements to prediction accuracy of QSAR models

Unsuccessful QSAR modeling is typically characterized by over-fitting of the trained model where the given ligand library is memorized rather than general patterns determined. Therefore, the primary goal of the modelling process is to achieve a high degree of generalizability when screening not yet seen ligand libraries. Best practices can be applied which comprise techniques such as cross-validation, consensus predictions, and molecular descriptor selection [29]. Cross-validation can be applied when the data set is subdivided into an independent, monitoring, and training partition. The advantage results because each data set fraction is part of an independent and monitoring dataset once in the life cycle of the cross-validation process [30]. This partitioning ensures less influence of extreme outlier data points in the independent dataset evaluation [31]. Another technique contributing to the robustness of QSAR modeling is the concept of consensus prediction [32]. Consensus measures take the prediction of different machine learning models and algorithms into account by taking the average of all predicted values or applying a jury approach where compounds are classified by a majority vote. Through this approach it is possible to compensate prediction errors of individual techniques. Furthermore, the selection of optimal molecular descriptors or features is another optimization process that can lead to a more robust QSAR model. Disproportionate numbers of descriptors or features can decrease the signal to noise ratio and thus reduce the predictive power of the model. Feature selection techniques can identify the most significant features while removing unnecessary descriptors to lower the dimensionality of the model. Ideally this reduces the involved degrees of freedom of the QSAR model and fosters its generalizability. Successful techniques involve feature selection including measures such as information gain [33], F-score [34], genetic

algorithm [35], swarm optimization [36] and whole descriptor group selection such as sequential feature forward selection and feature backward elimination [37].

High-throughput screening sets knowledge base for virtual screening in drug discovery

HTS is the process of testing a large number of diverse chemical structures against disease targets in order to identify new leads. Compared to traditional drug screening methods, HTS is characterized by its simplicity, rapidness, low cost, and high efficiency while taking ligand-target interactions as the principle dataset and leading to a higher information harvest. The screening attrition rate in the current drug discovery protocols suggests that one marketable drug emerges from approximately one million screened compounds. HTS can test hundreds of thousands of compounds per day, however, if fewer compounds could be tested without compromising the probability of success, the cost and time would be greatly reduced [38-40]. However, HTS often provides the necessary knowledge base for QSAR, and thus sets the foundation for virtual high-throughput screening (vHTS) in drug discovery.

Curating and identifying publicly and commercially available biological assay data sets

As part of LB-CADD, QSAR modeling shows encouraging results in identifying potential hit compounds that share a biological response [41-43]. This technology has strong potential to greatly reduce these costs in academic environments which are restricted by limited resources. However, LB-CADD methods and method development heavily depend on the availability of reliable HTS assay data sets and the determined relation of ligand structure and biological activity. It is a challenge to identify suitably refined data sets for QSAR modeling. Publically available compound repositories with biological assay data are of essence. Libraries such as PubChem [44, 45], ChemBL [46], or WOMBAT [47] provide access to such HTS experimental results which are typically associated with several hundred thousands of compounds are tested against different biological targets. The curation of HTS assay data into high-quality data sets for LB-CADD developments plays a pivotal role for LB-CADD method development. Best practices [29] for CADD emphasize the importance of data set curation. A research study [48] investigated public and commercial data repositories and revealed error rates ranging from 0.1 to 3.4% in respective databases.

With the increasing body of scientific studies in the field of CADD the importance of addressing data quality becomes inherently significant.

Virtual high-throughput screening expands chemical search space in comparison to traditional HTS

QSAR models employ machine learning techniques to perform virtual high-throughput screens such as identifying (virtual) molecules with predicted properties in a given value range. These molecules can be ordered or synthesized and experimentally assessed for the desired property. Machine learning is a key element in contemporary virtual high-throughput screening campaigns. Application of several types of ANNs for QSAR modeling were investigated by Burton et al. [49] including a review of difficulties and challenges. Winkler [50] presented an overview of ANNs and their limitations when accessing chemical structure data. SVMs were applied to mine HTS data for type I methionine amino peptidases inhibition in a study by Fang et al. [13]. A compound library of 43,736 organic small molecules was evaluated and reached a recovery rate of 50% when screening just 7% of the test set. Another study by Plewczynski et al. [51] evaluated the effectiveness of SVMs for a range of active compounds determined from the MDDR compound library. Sensitivity values of up to 80% and hit rates of up to 100% for specific targets were achieved. Sakiyama et al. [52] describe an approach combining SVMs and recursive partitioning by DTs to predict the metabolic stability of compounds. Stahura and Bajorath [53] investigated computational approaches, including SVMs, to complement HTS. Baurin et al. [54] surveyed various statistical and computational techniques, such as DTs in their 2D-QSAR models for COX-2 inhibition. This particular study was based on the NCI database with approximately 200,000 compounds. In general, machine learning has substantial impact on CADD through guiding HTS in drug discovery efforts.

Allosteric potentiators of mGlu₅ provide a novel approach for treatment of schizophrenia

As the primary excitatory neurotransmitter glutamate is responsible for rapid excitatory response of synapses in the mammalian central nervous system (CNS). The mediation of synaptic signals is regulated by a family of receptors categorized as ionotropic and metabotropic glutamate (mGlu) receptors.

Glutamate is the indigenous ligand that activates mGlu receptors, a part of the G-protein coupled receptors (GPCR) family. mGlu receptors are involved in various biochemical cascade reactions that are known to modify proteins and in turn participate in a large variety of CNS functions [55, 56] including memory and pain perception. The receptor subtype 5 (mGlu₅) is of particular interest. Recent studies provide strong support for the hypothesis that activators of mGlu₅ may provide a novel approach that could be useful in treatment of schizophrenia [57]. Unfortunately, it has been extremely difficult to develop highly selective agonists of mGlu₅ that have suitable properties for use as drugs. The glutamate binding site is highly conserved across mGlu receptor subtypes [55], making it difficult to develop highly selective glutamate-site ligands. Also, most glutamate site agonists are analogs of glutamate and do not possess pharmacokinetic properties or the ability to cross the blood-brain barrier, essentially crippling the use of these compounds as potential drugs. In addition, there are a number of problems associated with the use of direct acting agonists as drugs. These include adverse effects associated with excessive activation of the receptor, profound receptor desensitization, and loss of activity dependence of receptor activation. Recently, highly selective allosteric potentiators of these receptors, including mGlu₅, have been developed. These compounds do not activate the mGlu receptors directly but act at allosteric sites on the receptor to potentiate glutamate-induced receptor activation [58-60]. They provide an exciting advance in demonstrating the potential of this approach for developing novel therapeutic agents that increase activity of specific mGlu receptor subtypes. However, little is known about the precise domains involved in the action of different classes of mGlu receptor potentiators or the physiological impact of these compounds on mGlu receptor signaling in native systems. Thus, LB-CADD methods can play a pivotal role in guiding medicinal chemistry and drug discovery efforts to elucidate novel interrelations between new potentiators and mGlu receptors.

Novel inhibitors for *Plasmodium falciparum* have potential to diminish Malaria drug resistance

Malaria is an infectious disease carried from mosquitos to humans with an incidence of approximately 200 million cases a year. It is caused by Malaria parasites of the genus *Plasmodium* and is responsible for

more than 200 million cases with an estimated 660,000 human fatalities in 2010 [61]. This parasitic disease is clinically characterized by high fevers, anemia, and flu-like symptoms. Malaria causing parasites like *Plasmodium falciparum* release significant quantities of free heme as a digestive byproduct, a non-protein constituent of hemoglobin [62, 63]. A heme consists of an iron atom contained in the center of a heterocyclic porphyrin ring and is toxic to human cells. Parasites convert it into an insoluble crystalline form called hemozoin. Therefore, inhibition of hemozoin aggregation is a viable target against Malaria parasites. Novel antimalarial therapeutics have potential to diminish Malaria drug resistance. The scientific community now has access to an arsenal of thousands of compounds that are highly toxic to the parasite, but the process of developing these hits into robust lead compounds will be challenging. Identification of the molecular targets responsible for activity could aid in prioritization of these hits, while the lead optimization could be streamlined since the hit-target interactions could be directly studied which would save time and resources. While the experimental determination of targets is possible in some cases, it is again an expensive and time-consuming process. An affordable *in silico* screening approach would greatly benefit the efforts to discover new chemo types, and thus help diminish drug resistance of the Malaria parasite.

Predicting molecular properties through quantitative structure property relationships

In contrast to QSAR, Quantitative Structure Property Relationships (QSPR) seeks to establish a model correlating molecular properties to the molecular structure information of the molecule. Drug metabolism and pharmacokinetic (DMPK) properties of lead compounds include absorption, distribution, metabolism, excretion, and toxicity (ADMET) in humans and are a key criterion on whether a lead compound can be considered a viable drug candidate. Molecular properties such as the logarithm of the water/octanol partition coefficient (logP), intrinsic clearance, and plasma protein binding are of essence for bioavailability, duration, and intensity of therapeutics. Katritzky et al. [64] studied vapor pressures and aqueous solubility of 411 compounds spanning a diverse chemical space by applying QSPR models. The model achieved a correlation coefficient (R^2) of 0.949 and 0.879 for vapor pressure and for aqueous

solubility, respectively. The model was developed solely on the use of compound chemical structures. Together, the reliable prediction of vapor pressure and aqueous solubility allows for the prediction of water-air partition coefficients. A different study by Shen et al. [65] investigated the metabolic stability of drug candidates through QSPR modeling. A kNN approach was chosen and applied to predict properties for absorption, distribution, metabolism, and excretion. The biological assay data was provided by GlaxoSmithKline (GSK) through 631 diverse proprietary compounds. A final success rate of 83% accuracy is reported to correctly identify metabolically stable compounds in the human S9 homogenate. Overall, QSPR models can be trained to predicted DMPK properties which can be used as a filter in the drug discovery process to reduce the number of compounds necessary to evaluate.

Concepts of CADD are implemented in BioChemistryLibrary

All concepts of LB-CADD are implemented in the software suite BCL::ChemInfo, part of the BioChemistryLibrary (BCL). This software suite was developed in-house as an object-oriented library written in the programming language C++ to implement algorithms regarding small organic molecules and large molecular structures such as proteins. It contains more than 1,000 classes and approximately 500,000 lines of code. All machine learning algorithms and all molecular descriptors were implemented in this library. BCL::ChemInfo is a tailored method that streamlines data processing such as data set generation and cross-validation. The framework hosts a range of small molecule descriptors, descriptor selection strategies, and machine learning technologies. Speedups between 80 and 200 are achieved through OpenCL implementations of ANNs and SVMs used on graphics processing units (GPUs) hosted on an in-house CPU/GPU cluster. A command line interface is provided for easy access. Thus, no meta-language has to be learned like R or Matlab. It is an effective software package to prioritize potential hit compounds through virtual screening, and therefore guide drug discovery campaigns in medicinal chemistry and pharmacology. BCL::ChemInfo is freely available for academic use at www.meilerlab.org.

2. LIGAND-BASED VIRTUAL HIGH-THROUGHPUT SCREENING BENCHMARK AND METHOD DEVELOPMENT

This chapter is based on publication [3].

Introduction

The development of quantitative structure activity relationship (QSAR) models in ligand-based computer-aided drug design (LB-CADD) has shown practical value for *in silico* (virtual) high-throughput screening (HTS) to identify potential hit compounds, i.e., compounds that share a biological activity of interest [66]. However, the predictive power of QSAR models depends not only on molecular descriptors of chemical structure and mathematical models, but foremost on the size, quality, and composition of the training data. An increased need for QSAR analysis emerged with the advent of HTS in academic research [66]. Cost increases linearly with the number of compounds tested in the HTS experiment and the number of compounds physically available at one site is limited [67].

LB-CADD has the potential to reduce these costs in a resource-limited academic environment.

Public databases such as PubChem [44] contain biological activities for several hundred thousands of compounds tested against different biological targets [68]. Nevertheless, the number of compounds tested in a HTS experiment is typically at least two orders of magnitude smaller than the tens of millions of drug-like small molecules listed in PubChem. The space of possible drug-like molecules is even larger and estimated to be 10^{30} – 10^{60} [69]. LB-CADD has the potential to increase and diversify the chemical space tested.

From a methods perspective, increased public availability of large HTS data sets enables thorough benchmarking of existing LB-CADD methods and triggers the development of new cheminformatics tools; it will ultimately contribute to the fundamental understanding of protein-small molecule recognition and enhance the use of small molecule tools in biology. LB-CADD is particularly attractive in the

resource-limited environment of academia as it reduces the cost and increases quality of drug discovery and/or probe development for rare or neglected diseases.

Quantitative structure activity relationships relate chemical structure and biological activity

QSAR models seek to correlate the often complex non-linear relationship of chemical structure with the biological activity for a protein target [29, 70]. Dudek et al. [71] and Du et al. [72] provide an in-depth overview of current QSAR methods. Hansch et al. [8, 73] pioneered classical QSAR by investigating the biological activity of a set of compounds in relation to their corresponding physicochemical properties (hydrophobic, electronic, and steric effects) using linear regression models. Modern QSAR techniques employ fingerprints and 2D/3D descriptors coupled with machine learning methods [74, 75].

Molecular descriptors numerically encode chemical structure

The descriptors employed in this study (scalar, 2D/3D auto-correlation, radial distribution functions) are fragment-independent and transformation invariant. Fragment-independent molecular descriptors can encode the chemical structure of small molecules in a vector of constant length independent of compound size, composition, or position in space. Radial distribution functions [76] were successfully employed to study the A2A adenosine receptor agonist effect of 29 adenosine analogues [77]. A separate study focused on prediction of native receptor affinities of 38 vitamin D analogues [78] outperforming fragment-based molecular descriptors. Autocorrelation descriptors [79] were used to train machine learning models that predict CDK4/D inhibitory activity [80] and negative ionotropic activity of calcium entry blockers, among other applications [81].

Machine learning techniques have viable impact on the generation of QSAR models

Machine learning algorithms have shown exciting potential for developing QSAR models to predict biological activity data [17, 82, 83]. Machine learning methods recognize complex patterns and derive a model based on training data acquired from experimentally screened compound libraries. Subsequently,

large compound libraries can be screened virtually or *in silico* and ranked by predicted biological activity. This prioritizes a subset of compounds that is enriched for active molecules for acquisition or synthesis.

Mueller et al. [84] applied artificial neural networks (ANNs) to identify novel positive allosteric modulators for mGlu5, a G-protein coupled receptor (GPCR) involved in neurological disorders like schizophrenia in two separate experiments. QSAR models were trained based on a high throughput screen of approximately 144,000 compounds. These models were used to virtually screen commercially available compound libraries to prioritize the most potent compounds. The top ranked compounds were experimentally validated resulting in a significant hit rate increase of 28.2% compared to an initial experimental hit rate of 0.94%.

Golla et al. [85] successfully applied genetic and evolutionary algorithms to virtually screen for novel chemical penetration enhancers (CPEs) utilized through transdermal drug delivery. A set of 272 CPEs served as a pool for initial structures and QSAR model generation. A total of 4,834 molecules were generated by the genetic algorithm and 893 molecules were accepted having a score below a set threshold of 8. The study identified 18 novel CPEs that were experimentally evaluated for cytotoxicity and permeability, four of which express marginal to no toxicological effects.

In another study, Sun et al. [86] applied support vector machines in conjunction with 2D molecular descriptors to identify compounds involved in drug-induced phospholipidosis (PLD). PLD is implicated in intracellular accumulation of phospholipids and formation of concentric lamellar bodies. A set of 4,161 unique drug-like compounds from various small molecule libraries were evaluated in a quantitative HTS experiment. The resulting data was employed to train QSAR models. Using one third of the data as a training set, the final model achieved a prediction accuracy of 90% on the remaining two thirds of compounds.

Consensus of QSAR models has potential to improve prediction accuracy

Combination of different machine learning models can reduce the prediction error by compensating for the misclassification of any single predictor with the consensus of the remaining models [87]. Simmons et al. [88] compared several machine learning and chemometric methodologies used to develop ensemble classifiers on data sets derived from in vivo HTS campaigns. Model performance was compared using false negative and false positive error profiles. Ensemble classifiers constructed from methods like ANNs or DTs achieved true positive rates of over 80% in the top 1.4% of the ranked list with false positive rates between 5%–7%.

Svetnik et al. [89] introduced a procedure for building a sequence of predictive models using a Random Forest approach [90]. Each model is fitted to the gradient of a loss function in a stage-wise manner to analyze ten cheminformatics data sets. Results are comparable to those of other ensemble learning methods such as Bagging and Boosting and outperform regular decision trees with accuracy rates of over 80%.

On the other hand, Hewitt et al. [91] constructed QSAR consensus models based on Genetic Algorithms on smaller data sets for silastic membrane flux, toxicity of phenols to *Tetrahymena pyriformis*, acute toxicity to the fathead minnow and flash point. The data set sizes ranged from 250 to 605 compounds. The results suggest only marginal benefit for consensus models compared to a single model predictor.

Significance

The objective of this research is three-fold: (1) To compose a comprehensive benchmark set for ligand-based computer-aided drug discovery (LB-CADD)—i.e., cheminformatics. While PubChem is available since 2004, only now it grew to the size and quality needed to assemble a benchmark set of realistic HTS experiments, where actives have been confirmed experimentally, a wide range of relevant drug targets is spanned, and all data is available in the public domain. The data sets are carefully post-processed and made available so as to establish a benchmark for developing LB-CADD methods at

www.meilerlab.org/qsar_pubchem_benchmark_2012. (2) To substantiate anecdotal and isolated findings on best practices in LB-CADD. Cheminformatics studies have been published comparing different machine learning methods, testing different approaches to descriptor selection, and using consensus modeling approaches (see above). However, conclusions derived from these studies were often limited by small and/or unrealistic toy data sets or by studying only one of the aforementioned aspects in isolation. In result correlations between parameters remain uncertain, over-training and narrow application range on one class of target proteins are a major concern. The present study overcomes these limitations. (3) To introduce a cheminformatics framework BCL::ChemInfo that is freely available for non-commercial use. It exposes a variety of methods for molecular descriptor selection, and machine learning techniques including ANNs [22, 92, 93], SVMs with an extension for regression [18, 94-96], DTs [97-99], and KNNs [24, 25, 100]. It also enables consensus predictions from different machine learning models.

Results and discussion

Compilation of Validated PubChem HTS Screens Provides Benchmark Data Sets for Training QSAR Models

PubChem provides publically available libraries of small organic molecules that have been tested in a diverse set of HTS experiments. Primary screens often include many false-positive hit compounds which display a response in initial assay experiments but are inactive in confirmatory experiments. We focus on HTS experiments with a single well-defined biological target protein. With respect to the desired target, these hit compounds may include non-binders that act on a different component of the assay or binders that are non-specific to the target and recognize other biological molecules. To minimize the number of false-positive hit compounds we compiled nine data sets applying the following criteria: the HTS experiment must target one specific protein and contain a minimum of 150 confirmed active compounds. Further, we chose a diverse set of PubChem assays focused on pharmaceutically relevant small molecule protein targets such as GPCRs, ion channels, transporters, kinase inhibitors, and enzymes. All PubChem assays are identified by PubChem summary id (SAID) of the primary protein target and describe a collection of confirmatory screens for active compounds given by PubChem assay ids (AID). It proved

critical to go through a detailed manual verification of the HTS experiments performed and collate PubChem raw data to arrive at high-quality data sets. Complete data sets and their compilation protocols are provided in the Experimental Section. We propose that the data sets presented here can serve as a benchmark for further cheminformatics method development. An overview with statistics of all PubChem data sets can be found in Table 1. The data sets are made available at www.meilerlab.org/qsar_pubchem_benchmark_2012.

Table 1: Overview of PubChem biological assays and data set statistics.

Protein Target Class	Protein Target	PubChem Summary Assay ID SAID	Number Actives	Number Inactives	Hit Rate	Inactives -to-Actives Ratio
GPCR						
	Orexin1 Receptor	435008	234	218,071	0.11%	948
	M1 Muscarinic Receptor	1798	188	61,661	0.30%	327
	M1 Muscarinic Receptor	435034	448	61,407	0.73%	138
Ion Channel						
	Potassium Ion Channel Kir2.1	1843	172	301,473	0.06%	1,752
	KCNQ2 potassium channel	2258	213	302,351	0.07%	1,419
	Cav3 T-type Calcium Channels	463087	703	100,210	0.70%	143
Transporter						
	Choline Transporter	488997	252	302,246	0.08%	1,199
Kinase Inhibitor						
	Serine/Threonine Kinase 33	2689	172	319,821	0.05%	1,859
Enzyme						
	Tyrosyl-DNA Phosphodiesterase	485290	292	344,477	0.08%	1,179

Machine learning algorithms relate chemical structure to biological activity

Three supervised (ANN, SVM, DT) and one unsupervised (KN) machine learning approaches were evaluated to predict biological activity on confirmatory HTS assay data. Machine learning methods can

describe non-linear relations and recognize patterns within large sets of numerical descriptors. Trained models can adapt to complex interrelations and are capable of detecting even small signals at high noise levels. Likewise, non-linear methods are applied when no simple mathematical model can be assumed, many influencing factors interact, and the experimental uncertainty is high.

Quality measures assess the predictive power of machine learning algorithms

The low ratio of active:inactive compounds in HTS data sets (typically 1:100–1:1000) leaves unbiased quality measures (e.g., classification accuracy) inappropriate for comparing results across different data sets. To facilitate meaningful comparisons of results across data sets, the integral of true-negative-rate (specificity, y-axis) and true-positive-rate (sensitivity, x-axis) (TNR-TPR) was chosen as an objective function for QSAR model training. It represents a 90° clock-wise rotation of the traditional receiver operating characteristic (ROC) curve [101]. The integral is identical to the well-known area under the curve (AUC) value. The TNR-TPR plot shows the accuracies of a model at predicting actives and inactives on separate axes, and is thus independent of the active:inactive ratio. It thereby preserves the key advantage of the ROC curve. However, it also eliminates a key disadvantage of the ROC curve: after virtual screening only a small fraction of compounds—the ones with predicted high activity—will be considered. How many actives are among these compounds depends on the very initial slope of the ROC curve which is difficult to measure with an integral as the optimal cutoff value on the x-axis tend to be very small (for example 10^{-3}) and data set dependent, i.e., it needs to be adjusted for optimal performance. In the past the x-axis has therefore been often plotted on a logarithmic scale. For TNR-TPR plots the integration is performed from 0 to the desired TPR value—i.e., the fraction of actives recognized as such, for example 25% or 50%. The integral instead of the slope is now the determinant of model quality. In contrast the slope analyzed for ROC curves at a single point the integral computed for TNR-TPR curves presents a more robust measure for model quality (see Experimental Section). To facilitate comparison we report enrichment (ENR) as additional quality measure—i.e., what is the ratio of active

compounds after virtual screening compared to the initial HTS experiment. Enrichment correlates with the slope of the ROC curve.

QSAR model quality depends critically on the selection of optimal descriptor set

To systematically add the most significant descriptor elements for signal to noise increase, three descriptor selection methods were employed: Information Gain (IG), F-Score (FS), and Sequential Forward Feature Selection (SFFS). A total of 60 numerical descriptor groups were available for this study with a total of 1,284 descriptor values (see supplementary materials Table S1). All three methods were used separately for each machine learning technique and PubChem data set.

To cope with the computational expense associated with each descriptor selection technique, reduced data sets were created for the training and monitoring data sets containing always all available active compounds. Sets of 30,000 and 10,000 inactive compounds for the training and monitoring partitions, respectively, were chosen randomly from each HTS data set. The independent data set was not altered. The optimal descriptor set is defined by the largest average integral of the TNR-TPR curve over all cross-validation experiments. The descriptor selection process aims to identify the combination of properties and encoding functions that describe the structural features of the pharmacophore best and hence yield the best QSAR model. The selection process allows for models with fewer degrees of freedom and therefore reduces training time, limits number of data points needed for training, and improves signal to noise. For every descriptor selection method, machine learning algorithm, and PubChem data set a comparison of the integral beneath the TNR-TPR curve was evaluated by assessing 10×9 -fold cross-validated models using the optimal descriptor set shown in Table 2.

Table 2: Overview of descriptor selection results.

Descriptor selection results for Information Gain (IG), F-Score (FS), and Sequential Feature Forward Selection (SFFS) applied to each machine learning technique paired with each PubChem HTS data set. Results for the integral of the TNR-TPR curve are presented. The mean and standard deviation for each SAID (row) and each descriptor selection approach (column) is given.

PubChem SAID	ANN			SVM			DT			KN			Mean (Stdev)
	IG	FS	SF FS	IG	FS	SF FS	IG	FS	SF FS	IG	FS	SF FS	
435008	0.79	0.81	0.80	0.84	0.84	0.85	0.77	0.77	0.77	0.77	0.77	0.77	0.80 (0.03)
1798	0.68	0.68	0.64	0.74	0.73	0.76	0.72	0.68	0.52	0.70	0.72	0.68	0.69 (0.06)
435034	0.79	0.80	0.80	0.83	0.82	0.85	0.74	0.74	0.50	0.75	0.76	0.78	0.76 (0.09)
2258	0.80	0.80	0.83	0.84	0.84	0.85	0.78	0.75	0.77	0.75	0.76	0.79	0.80 (0.04)
1843	0.91	0.90	0.92	0.92	0.91	0.92	0.86	0.84	0.83	0.86	0.83	0.86	0.88 (0.04)
463087	0.84	0.86	0.86	0.88	0.89	0.89	0.82	0.81	0.82	0.75	0.77	0.81	0.83 (0.05)
488997	0.76	0.75	0.75	0.79	0.81	0.82	0.77	0.74	0.75	0.74	0.75	0.76	0.77 (0.03)
2689	0.92	0.92	0.92	0.92	0.93	0.93	0.88	0.86	0.86	0.88	0.86	0.85	0.89 (0.03)
485290	0.83	0.84	0.85	0.86	0.86	0.86	0.82	0.84	0.76	0.80	0.80	0.75	0.82 (0.04)
Mean (Stdev)	0.81 (0.07)	0.82 (0.07)	0.82 (0.09)	0.85 (0.06)	0.85 (0.06)	0.86 (0.05)	0.79 (0.05)	0.78 (0.06)	0.73 (0.13)	0.78 (0.06)	0.78 (0.04)	0.78 (0.05)	

The mean value of the TNR-TPR integral comparing the performance of machine learning algorithms across different descriptor selection methods ranged from 0.73–0.79 (DT), 0.78 (KN), to 0.81–0.82 (ANN) and 0.85–0.86 (SVM). These results provide a clear distinction of prediction performance between individual single predictors. The mean performance of a cross-validated QSAR model considering each PubChem data set individually ranged from 0.69 (SAID 1798) to 0.89 (SAID 2698); standard deviations ranged from 0.03 to 0.09. ANN and SVM typically outperform DT and KN with mean integral values above the baseline independent of the chosen data set.

Consensus prediction of machine learning techniques increases prediction accuracy

A consensus prediction was obtained by averaging the output of several machine learning techniques (see Experimental). For every data set, all possible combinations of machine learning methods were assessed

using the previously determined optimal descriptor set for each machine learning method (see Table 3). We ranked QSAR model performance using the achieved ENR value. The best single predictor is listed at the end of each ranking if not present among the top three consensus predictors.

Consensus models consistently outperform QSAR models that rely on a single machine learning method. Misclassifications of single predictors are generally extenuated and therefore consensus predictors achieve increased prediction accuracy. Normalizing the ENR increase between the best consensus predictor and best single predictor by the inactives-to-actives ratio reveals differences ranging from -0.21% (SAID 438005, IG) to 4.75% (SAID2689, SFFS). In all but a single case the consensus predictor outcompetes individual predictors suggesting that consensus prediction provides a benefit although it is small. The overall ENR increase of consensus predictors compared to the theoretical maximum ENR appears to be marginal, though significant when compared to the hit rate of the HTS experiment (see Table 1). Conflicting findings from previous studies [87-91] can be attributed to a limited benchmark.

Table 3: Overview of consensus benchmark results for all PubChem datasets

SAID	FS	#r	INT	ENR	IG	#r	INT	ENR	SFFS	#r	INT	ENR
435008	ANN DT KN SVM	1	0.249	27	SVM	1	0.245	24	DT SVM	1	0.247	40
	SVM	2	0.245	26	DT SVM	2	0.245	22	SVM	2	0.246	39
	DT SVM	3	0.245	26	KN SVM	3	0.245	22	ANN DT SVM	3	0.249	27
	Diff	-	-	0.11%	Diff	-	-	-0.21%	Diff	-	-	0.11%
1798	ANN DT KN SVM	1	0.240	10	ANN DT KN SVM	1	0.241	15	ANN DT KN SVM	1	0.243	7
	ANN DT SVM	2	0.240	9	ANN KN SVM	2	0.241	10	ANN KN SVM	2	0.243	7
	SVM	6	0.240	8	SVM	8	0.242	7	SVM	8	0.244	5
	Diff	-	-	0.61%	Diff	-	-	2.45%	Diff	-	-	0.61%
435034	ANN SVM	1	0.246	18	ANN SVM	1	0.245	17	ANN SVM	1	0.246	18
	ANN DT SVM	2	0.246	17	ANN DT SVM	2	0.245	16	ANN DT SVM	2	0.246	18
	SVM	7	0.245	14	SVM	4	0.246	16	SVM	3	0.246	17
	Diff	-	-	2.90%	Diff	-	-	0.72%	Diff	-	-	0.72%
1843	ANN DT KN	1	0.249	54	ANN DT SVM	1	0.250	68	ANN DT KN SVM	1	0.250	50
	DT KN SVM	2	0.249	45	ANN DT KN SVM	2	0.250	67	ANN DT SVM	2	0.250	45
	SVM	11	0.249	32	SVM	10	0.250	45	ANN	11	0.250	30
	Diff	-	-	1.26%	Diff	-	-	1.31%	Diff	-	-	1.14%
2258	ANN DT	1	0.241	28	ANN DT	1	0.241	34	ANN DT KN	1	0.244	65
	ANN DT SVM	2	0.246	26	ANN DT KN SVM	2	0.246	33	DT KN SVM	2	0.249	49
	SVM	8	0.246	18	DT	6	0.238	23	SVM	10	0.249	28
	Diff	-	-	0.70%	Diff	-	-	0.78%	Diff	-	-	2.61%

The top two ranking (#r) consensus QSAR models in Table 3 are presented for each PubChem data set (SAID) and descriptor selection techniques: F-Score (FS), Information gain (IG), and sequential feature forward selection (SFFS). Every model is evaluated by the integral of the TNR-TPR curve with a TPR rate of 0.00 to 0.25 (INT). The ranking is ordered by Enrichment (ENR). The best single predictor is shown for comparison. The ENR difference of the best consensus predictor compared to the best single predictor normalized by the inactives-to-actives ratio is given (Diff).

Table 3 continued.

SAID	FS	#r	INT	ENR	IG	#r	INT	ENR	SFFS	#r	INT	ENR
463087	ANN SVM	1	0.250	23	ANN DT KN SVM	1	0.250	19	ANN KN SVM	1	0.250	29
	SVM	2	0.250	22	ANN DT SVM	2	0.250	18	ANN DT KN SVM	2	0.250	28
	DT SVM	4	0.250	21	SVM	8	0.250	17	SVM	8	0.250	24
	Diff	-	-	0.70%	Diff	-	-	1.40%	Diff	-	-	3.50%
2689	ANN DT KN SVM	1	0.248	74	ANN DT KN SVM	1	0.248	58	ANN DT SVM	1	0.249	101
	ANN DT SVM	2	0.248	63	ANN DT SVM	2	0.248	54	ANN DT KN SVM	2	0.249	91
	ANN	100	0.250	42	SVM	100	0.248	41	ANN	100	0.248	44
	Diff	-	-	2.67%	Diff	-	-	1.42%	Diff	-	-	4.75%
488997	ANN DT SVM	1	0.246	20	DT KN SVM	1	0.244	14	ANN DT KN SVM	1	0.243	49
	ANN DT KN SVM	2	0.247	19	ANN DT KN	2	0.241	13	ANN KN SVM	2	0.243	44
	SVM	6	0.245	15	DT	7	0.242	12	SVM	110	0.244	31
	Diff	-	-	0.27%	Diff	-	-	0.11%	Diff	-	-	0.97%
485290	ANN DT KN	1	0.241	64	DT KN SVM	1	0.245	71	ANN SVM	1	0.244	30
	DT KN SVM	2	0.245	58	ANN DT KN SVM	2	0.246	60	ANN DT SVM	2	0.244	28
	SVM	110	0.244	38	SVM	120	0.245	36	SVM	4	0.244	26
	Diff	-	-	2.22%	Diff	-	-	2.96%	Diff	-	-	0.28%

Experimental

Determination of confirmatory high-throughput screening data sets for diverse protein targets

Publicly available libraries of small organic molecules from a diverse set of HTS experiments were obtained from PubChem. The following listing of PubChem assays identifies the PubChem summary id (SAID) of the primary protein target and describes the determination of active compounds from

confirmatory screens given by PubChem assay ids (AID). The inactive compounds are taken from the corresponding primary assay.

GPCR: Antagonist of the Orexin 1 Receptor (SAID 435008)

The GPCR Orexin 1 plays a role in behavioral plasticity, the sleep-wake cycle, and gastric acid secretion [102, 103]. Three primary screens, AID 485270, AID 463079, AID 434989, were conducted to identify antagonists of Orexin 1 receptor. AID 485270 is a FRET-based cell-based assay [104] identifying compounds that inhibit Orexin 1 receptor activity. AID 463079 is a cell-based assay using a parental CHO cell line identifying compounds that non-selectively inhibit Gq signaling. Here, compounds are tested for inhibition of Gq activity using the parental CHO cell line without transfection of the GPCR. AID 434989 is a fluorescence-based cell-based assay identifying compounds with inhibitory activity of the Orexin 1 receptor. These compounds are dispensed onto CHO cells with transfected human Orexin 1 receptors to gauge calcium mobilization by a fluorescent indicator dye. Inhibitors revealed by the primary screen AID 485270 were confirmed by the counter screen AID 492964 through a Homogeneous Time Resolved Fluorescence (HTRF)-based cell-based assay. Further, resulting inhibitors from assay AID 492964 were investigated by the counter screen AID 493232 that tested for non-selectivity due to inhibition of Gq activity. It applied a HTRF-based cell-based assay to identify antagonists of the parental CHO-K1 cell line. Subsequently, the validation assay AID504701 identified compounds being active in primary screen AID 483270, confirmed in assay AID 493964, but inactive in counter screen AID 493232 to exclude compounds with non-selectivity due to inhibition of Gq activity. AID504701 applied an HTRF-based cell-based dose response assay to identify antagonists of the Orexin 1 receptor. Another validation screen, AID 492965, confirmed compounds, active in AID 434989 and inactive in primary screen AID 463079, being non-selective inhibitors of Gq signaling. The applied assay was a fluorescence-based cell-based HTS confirmation assay. A second primary assay, AID 434989, screened for agonists of the Orexin 1 receptor with validation of inhibitory activity by AID 492963. AID 492963 used a fluorescence-based cell-based HTS confirmation assay to identify antagonists of the Orexin 1 receptor. A

more specific assay, AID504699, identified compounds that are active in AID 434989 and AID 492963 but being inactive against the parental cell line tested in a third primary screen AID 463079. AID504699 applied a fluorescence-based cell-based HTS confirmation assay to identify antagonists of the Orexin 1 receptor. Combining the active compounds of the most refined assays, AID504701 and AID504699, resulted in a total of 234 active compounds excluding an overlap of 155 molecules.

GPCR: Allosteric modulators of M1 Muscarinic Receptor: Agonist (SAID 1798)

The Gq-coupled GPCR M1 Muscarinic Receptor [105-108] is a seven-transmembrane domain receptors whose modulation has significant impact in treatment of cognitive degeneration associated with Alzheimer's disease and schizophrenia. The same set of compounds was screened for positive (AID626) and negative (AID628) allosteric modulation of the M1 Muscarinic receptor. Agonistic modulators of M1 Muscarinic receptor were confirmed by screen AID 1488 applying a cell-based fluorometric calcium assay. A second counter screen AID 1741, using the same assay settings as AID 1488, evaluated these compounds for cross-activity with M4 muscarinic receptor. The final set of selective positive allosteric modulators of M1 was obtained by removing compounds active in AID 1741 from the compounds active in AID 1488 resulting in 188 compounds.

GPCR: Allosteric modulators of M1 Muscarinic Receptor: Antagonist (SAID 435034)

Negative modulators of M1 muscarinic receptors (AID628) [109, 110] were confirmed by screen AID677 through a cell-based fluorometric calcium assay. AID859 confirmed activity on rat M1 receptor. The counter screen AID860 removed non-selective compounds being active also at the rat M4 receptor. AID859 and AID860 employed the same assay type as AID677 and AID628. To remove the non-selective actives having a different target than the rat and human M1 receptor, the final set of active compounds was obtained by subtracting active compounds of AID860 from those in AID677, resulting in 448 total active compounds.

Ion Channel: Potentiators of KCNQ2 potassium channel (SAID 2258)

Voltage-gated potassium channels, like KCNQ2 [111, 112], have important neuronal functionality in excitement and resting states of cells. This target institutes a new avenue for drugs attempting to treat cancer, autoimmune diseases, and metabolic, neurological, and cardiovascular disorders. The primary screen AID 2239 identified potentiators of KCNQ2 potassium channel through measurements of intracellular thallium, gauged by the intensity of a thallium-sensitive fluorescent dye. A confirmatory screen AID 2287 validated active compounds to be potentiators. Counter screens identified false positive compounds showing response for CHO-K1 cell activity (AID 2282), non-specific effects on KCNQ1 (AID 2283) and response in KCNQ2-W236L-CHO cells (AID 2558). All confirmatory and counter screens applied the same experimental conditions as in the primary screen. The final set of 213 active compounds was acquired by removing the active compounds of AID 2282, AID 2283 and AID 2558 from the confirmatory screen active set of compounds (AID 2287).

Ion Channel: Identification of compounds that inhibit Inward-Rectifying Potassium Ion Channel Kir2.1 (SAID 1843)

The Kir2.1 inward-rectifier potassium ion channel is recognized as a target in the treatment of cardiovascular, neurological, renal and metabolic disorders [113-115]. The primary assay AID 1672 identified inhibitors for the inward-rectifying potassium ion channel Kir2.1. The assay uses a HEK293 cell line with stably expressed Kir2.1 channels where test compounds are gauged by intracellular thallium through thallium-sensitive fluorescent dye. The validation screens AID 2032 and AID 463252, both confirmed active compounds from the primary screen showing inhibition of Kir2.1. While AID 2032 used the same assay experiment as the primary screen, AID 463252 applied an automated electrophysiology assay for Kir2.1. The counter screens AID 2105, AID 2345, AID 2236, and AID 2329 identified active compounds exhibiting non-specific binding effects against Kir2.1. AID 2105 is a counter screen to the primary screen and evaluated active compounds against their non-specific effects on parental HEK293 cells of Kir2.1-HEK293 cells. AID 2236 tested compounds identified in the primary screen assay for effects on hERG CHO cells. AID 2345 assess compounds identified as active in an independent primary

screen assay (PubChem AID 2239) for non-specific effects on the Kir2.1 stably expressed HEK293 cells as well as on KCNQ2 potassium channels.

The final set of 172 active compounds was assembled by subtracting the actives in AID 2105, AID 2345, AID 2236, and AID 2329 from the molecules found active in both, AID 2032 and AID 463252.

Ion Channel: Inhibitors of the of Cav3 T-type Calcium Channels (SAID 463087)

The transient-type (T-type) calcium channel containing one of three α 1 subunits (Cav3) is part of the voltage-gated potassium channel family and has suggested involvement in epileptics and pulmonary hypertension [116-119]. The primary screen AID 449739 identified inhibitors of Cav3 T-type calcium channels measuring calcium fluorescence change in a Cav3.2 expressing cell line. AID 489005 is a counter screen validating active compounds of the primary screen using the same assay. Four follow-up screens were performed to confirm inhibitory effects on smaller sets of compounds involving AID 493021, AID 493022, AID 493023, and AID 493041. All were confirmatory and tested dose-response through 11-point 3-fold experiments using the same assay conditions as the primary screen.

The final set of 703 active compounds was acquired by subtracting the inactive compounds of the latter follow-up screens from the actives in the validation screen AID 498005. Taking just the actives of the four follow-up screens would have violated the established benchmark data set requirements.

Transporter: Inhibitors of the Choline Transporter (CHT, SAID 488997)

Choline has many physiological functions throughout the body that are dependent on its available local supply [120, 121]. Its transport is required for cellular membrane construction and is the rate-limiting step for acetylcholine production. CHT is suggested a drug target involved in Alzheimer's disease. The primary screen AID 488975 identified inhibitors of CHT. The counter screen AID 493221 is a validation screen to confirm the active compounds that inhibit CHT. It uses a choline-induced membrane potential assay measuring choline coupled sodium flow through CHT. Further, two additional validation screens reaffirmed activity of these compounds with 5-point concentration response curve (CRC) (AID504840)

and 10-point CRC (AID588401) experiments. The screen AID 493222 evaluated remaining active compounds for non-specific activity in parental HEK293 cells. Finally, the reconfirmation screen AID602208 tested a selected set of compounds for 3H choline uptake.

The final set of 254 active compounds was determined by the overlap of active compounds in screens AID 493221, AID504840, and AID588401 subtracting any non-specific hits from AID 49322 and all inactive compounds in the re-confirmation screen AID602208.

Kinase Inhibitor: Inhibitors of Serine/Threonine Kinase 33 (STK33, SAID 2689)

The serine/threonine kinase, STK33, has been shown to be required for the survival and proliferation of mutant KRAS-dependent cells involved in cancer [122]. The primary screen AID 2661 identified inhibitors of STK33 through preincubation of purified STK33 Kinase with potential inhibitors and kinase activity is measured through luminescent signal strength. The counter screen AID 2821 reaffirmed active compounds from AID 2661 using the same experimental conditions as in the primary screen. AID504583 tested a subset of compounds for STK33 selectivity by measuring Protein Kinase A inhibition. Taken the actives in AID 2821 subtracted by the actives from screen AID504583 resulted in the final set of 172 active compounds.

Enzyme: Inhibitors of Tyrosyl-DNA Phosphodiesterase 1 (TDP1, SAID 485290)

The inhibition of Human tyrosyl-DNA phosphodiesterase 1 (TDP1) has the potential to enhance anticancer activity of DNA topoisomerase I inhibitors [122-124]. The primary screen AID 485290 identified inhibitors of TDP1. The counter screen AID 489007 was used as a confirmation of the previously identified actives. AID 489007 used the Alpha Screen detection method [124] measuring the intensity of an enzyme cleavage reaction. The final set of 292 actives contains all compounds labeled as active in the counter screen AID 489007.

Numerical representation of biological data distinguishes active from inactive compounds

The half maximal inhibitory and effective concentrations, IC_{50} and EC_{50} values, of active compounds from the HTS data ranged from 0.1 μM to 25 μM . Biological activity was not reported for every active compound, thus all active compounds without an assigned IC_{50} or EC_{50} value were categorized as actives with a representative value of 1 μM chosen from the actives concentration range. All inactive compounds were set to a biological activity value of 1 mM. Models were trained on $pIC_{50} = -\log_{10}(IC_{50}/1 \text{ M})$, which ranges from 3 (for inactive molecules) to 4.6 to 7 (for molecules with $IC_{50} = 25$ to 0.1 μM). The same method was applied for pEC_{50} values. This procedure ensures that compounds without determined IC_{50}/EC_{50} can be used for training while at the same time information on differential activity is leveraged when available. In our hands this procedure is superior to a pure binary classification in active/inactive (data not shown).

Numerical description of molecules for QSAR model development

A total of 1,284 numerical descriptors in 60 categories were implemented in this study (see supplementary materials Table S1). The 60 categories contain scalar descriptors such as molecular weight, number of hydrogen bond donors, -acceptors, octanol / water partition coefficient, total charge, and topological polar surface area. Nine additional chemical properties were computed for every atom including atom identities, σ -, π -, and total charges, σ -, π -, and lone pair electronegativities, effective atom polarizabilities, and VC2003 atom charges [125]. Three encoding functions (2D auto-correlation, 3D auto-correlation, radial distribution function) are paired with each of the chemical properties to yield 27 fingerprints [77, 80]. In addition, each fingerprint is computed a second time applying van der Waals surface area as a weight factor. 3D conformations for all molecules were calculated with CORINA [126].

Monitoring data set is used for early termination of training process

The oversampled data set had 80% of the data points employed in the actual training process. The number of training iterations was limited through early termination to counter “overfitting” of the machine learning model to the training data. A monitoring data set consisting of 10% of the data points was used

to optimize all training parameters of the machine learning methods and to invoke early termination. The final 10% of the data points are set aside as an independent data set. It is not employed in the training process, but was evaluate by the final model. There was no overlap of compounds between training, monitoring, and independent data sets.

The integral of the true-negative-rate–true-positive-rate curve is a viable quality measure for QSAR models

Similarly to a traditional receiver operating characteristic (ROC) curve, QSAR models were evaluated by means of a true-negative-rate - true-positive-rate (TNR-TPR) curve. It resembles a clockwise rotated ROC curve plotting the rate of true negatives $TNR = TN/N = 1 - FR/N = TN/(FP + TN)$ (or specificity) versus the rate of true positives $TPR = TP/P = TP/(FN+TP)$, also known as sensitivity. The diagonal represents performance of a random predictor and has an integral (area under the curve, AUC) of 0.5. The QSAR model progressively improves as the integral increases. For virtual screening where only a small fraction of a screened compound library will be tested experimentally, performance at high true positive rates (or low false positive rates) is most critical (see below).

Enrichment measures ratio of fraction of active compounds predicted above actives rate

QSAR models are most frequently applied in virtual screening experiments: activities are predicted for a large compound library (e.g., $\sim 10^5$). Compounds are ranked by predicted activities to select a small fraction of compounds for experimental testing (e.g., 1% or $\sim 10^3$). In this scenario it is important that the 1,000 compounds predicted as most active are actually active while the ranking of the other 99,000 compounds is of a lesser concern. This property of the QSAR model is not well reflected in the global AUC value as it only depends on the initial integral of the TNR-TPR curve. It is better analyzed through computation of enrichment:

$$ENR = \frac{\frac{TP}{FP+TP}}{\frac{P}{P+N}} \quad (1)$$

The value represents the factor by which the fraction of active compounds was increased through virtual screening above the background observed in the original HTS experiment.

Orthogonal supervised and unsupervised machine learning algorithms seek optimal biological activity predictions

A set of four supervised and unsupervised machine learning algorithms were implemented in this study. The first supervised algorithm is the Artificial Neural Networks (ANN). The utility of ANNs for classification is well-known in chemistry and biology [28, 82, 83, 127]. Their architectural arrangement resembles the network structure of neurons. Layers of neurons are connected by weighted edges w_{ji} . The input data x_i are summed according to their weights, an activation function is applied, and its output used as the input to neurons of the next layer. Simple propagation [92] was chosen as the weight update algorithm during the training process. The parameters (learning rate) and α (momentum) were optimized prior to descriptor selection and again with the optimized descriptor set.

Support Vector Machine (SVM) learning with extension for regression estimation [18] represents a supervised machine learning approach successfully applied in the past [82, 128-130]. The core principles lay in linear functions defined in high-dimensional hyperspace [21], risk minimization according to Vapnik's ν -intensive loss function, and structural risk minimization [96] of a risk function consisting of the empirical error and the regularized term. SVMs were trained using an initial penalty parameter C and kernel parameter γ of 1 and 0.1 respectively, during the descriptor optimization process. Upon identification of the optimal descriptor set, C and γ were optimized in a grid search approach for every data set.

The decision tree (DT) learning algorithm [88, 89, 98] determines sets of rules to partition a given training data set. A partitioning algorithm gauges each successive split into subset or decision nodes with increased purity of one small molecule category. The splitting criterion is ascertained by the Gini coefficient [131]. The resulting model is a sequence of decisions involving single predictor variables that

classify a given feature. The DT implementation in BCL::ChemInfo ranks outcomes based on the percent of actives that were mapped into that node during training.

The Kohonen network (KN) represents an unsupervised learning algorithm [24, 25, 100]. The KN clusters similar inputs into nodes on a spatial grid, thus forming a reduced-dimensional representation of the problem space. Each node represents a cluster of similar compounds based on a Gaussian neighbor kernel distance measure.

Cross-validation ascertains robustness of QSAR models

The active and inactive data sets are divided into ten equal-sized partitions. The first partition is specified as the independent data set which is constant during cross-validation. Of the remaining nine partitions a second partition is selected as the monitoring data set. The remaining eight subsets constitute the training data set. A different monitoring data set is chosen systematically for each iteration of the cross-validation. In a set of ten data partitions each of those ten partitions can be assigned as independent data set leaving nine possibilities of assigning one remaining data partition as the monitoring data set. This results in $10 \times 9 = 90$ possible model training configurations. All final models trained using the optimized descriptor sets in this study are 10×9 -fold cross-validated. This procedure still ensures that every molecule in the data set was part of an independent data partition at least once during cross-validation. Data sets for ANNs and SVMs were balanced by oversampling actives, while decision trees and KNs required no oversampling.

To reduce the computational burden, all descriptor selection schemes use a $5 \times 1 = 5$ fold cross-validation set up, where the monitoring data partition is systematically incremented but only one independent data set configuration is evaluated.

Selection of an optimized descriptor set guides QSAR model training

To reduce the total number of inputs to machine learning algorithms, it is advantageous to remove obsolete descriptors in order to minimize the number of degrees of freedom that need to be determined. Further, noise is reduced while the ratio of data points versus degrees of freedom increases. The

determination of an optimal set of descriptors for each data set was evaluated by various selection methods such as Information gain [33], F-Score [34], and Sequential Forward Feature selection [132].

Information gain (IG) and F-Score (FS) score every descriptor column in the data set by statistical metrics that consider the actives/inactives composition. These scores can be compared: higher values indicate a higher discriminating power between active and inactive data points for a particular descriptor column. A particular advantage of these metrics over SFFS is that relatively few models need to be trained, because the scores are independent from model training.

IG measures the change of information entropy from the overall compound distribution of actives and inactives in one descriptor column compared to the entropy in each descriptor category itself. A higher information gain value of a descriptor column indicates higher discriminating information content.

$$\text{Information gain: } IG(x) = -\sum_{i=1}^m x_i \log_{10} x_i \quad (2)$$

The variable represents the i -th feature of the combined active and inactive data sets. FS considers the mean and standard deviation of each descriptor column across active and inactive compounds

$$\text{F-Score: } F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (3)$$

where are the average of the i -th feature of the whole, active, and inactive data sets, respectively; is the i -th feature of the k -th active instance, and is the i -th feature of the k -th inactive instance.

SFFS evaluates the objective function of trained models directly to arrive at an optimal descriptor set. This approach is a deterministic greedy search algorithm over all descriptor groups (see supplementary materials Table S1). Each round adds a single descriptor group to the descriptor set (initially, the empty set) selected in the previous round. Descriptor sets for the current round are then formed by adding each candidate descriptor group to the descriptor set selected in the previous round. Descriptors already present in the best descriptor group are ignored when creating the descriptor sets for a given round. Five-fold

cross-validated models are trained followed by the evaluation of respective objective functions. The average objective function result is computed for each cross-validated model, and the descriptor set corresponding to the top performing models is selected as the best descriptor set for this round. This process is repeated until all features are selected or early terminated if no improved was determined for ten consecutive rounds. Finally, the best descriptor combination is chosen from the best performing model.

Consensus predictions seeks improved accuracies of trained QSAR models

The combination of different machine learning model predictions can reduce the overall prediction error by compensating for misclassification of a single predictor with the consensus of the remaining models [87]. Here, we evaluate the overall accuracy of all trained QSAR models by calculating average consensus of all predicted pIC₅₀ or pEC₅₀ values given in an independent data set:

$$\text{Consensus :: average} = \frac{1}{N} \sum_{i=1}^N pIC_{50} \text{ or } \frac{1}{N} \sum_{i=1}^N pEC_{50} \quad (4)$$

If the predicted pIC₅₀ or pEC₅₀ value is at or above a given cutoff the predicted activity of the molecule is evaluated as active. Consensus, prediction was performed over all cross-validated QSAR models, five cross-validation models for every of the four machine learning methods.

Implementation

All machine learning algorithms and all molecular descriptors were implemented in the BioChemistryLibrary (BCL). This software suite was developed in-house as an object-oriented library written in the programming language C++. It contains more than 1,000 classes and approximately 500,000 lines of code. This library is the basis for BCL::ChemInfo and other modeling algorithms regarding small organic molecules and large molecular structures such as proteins. BCL::ChemInfo is a tailored method that streamlines data processing such as data set generation and cross-validation. The framework hosts a range of small molecule descriptors, descriptor selection strategies, and machine

learning technologies. Speedups between 80 and 200 are achieved through OpenCL implementations of ANNs and SVMs used on graphics processing units (GPUs) hosted on an in-house CPU/GPU cluster. A command line interface is provided for easy access. Thus, no meta-language has to be learned like R or Matlab. BCL::ChemInfo is freely available for non-commercial use at www.meilerlab.org/bclcommons.

Conclusions

In this study, nine large data sets were assembled originating from realistic HTS experiments for a range of common drug target proteins including GPCRs, ion channels, transporters, kinase inhibitors, and enzymes. All data was drawn from the public domain through PubChem but carefully post-processed to include only confirmed active compounds. These data sets provide a foundation for developing and testing methods in LB-CADD. We further introduce a comprehensive framework for LB-CADD termed BCL::ChemInfo that is freely available for non-commercial use and exposes orthogonal machine learning approaches including ANNs, SVMs, DTs, and KNs. We confirm that the quality of QSAR models depends critically on selection of optimal molecular descriptors, composition of the training data, and the machine learning method itself.

Further optimization was achieved by combining different machine learning methods into a consensus prediction to reduce false positives of each individual method. Theoretical enrichments ranging from 15 to 101 for a TPR cutoff of 25% are observed. The overall enrichment improvement normalized by maximal possible enrichment of consensus predictors compared to single predictors is up to 4.75%. We derive a ‘TNR-TPR curve’ from the common ROC analysis to better evaluate the quality of QSAR models at high TPR/FPR ratios.

LB-CADD or ‘cheminformatics’ is one strategy to reduce costs and increase size of the chemical space in resource-limited academic probe development efforts or drug discovery campaigns that target orphan or neglected diseases. This study shows that QSAR models prioritize compounds *in silico* thereby limiting

the cost of HTS and hit-to-lead optimization. The availability of HTS data through PubChem allows for a comprehensive comparison of QSAR models, molecular descriptor selection, and training strategies.

3. IDENTIFICATION OF PATHWAY SPECIFIC INHIBITORS FOR B-HEMATIN CRYSTALLIZATION IN PLASMODIUM FALCIPARUM INVOLVED IN MALARIA

Introduction

(This section was kindly provided by Rebecca D. Sandlin)

The malaria eradication campaigns of the mid 1900s dramatically lowered infections worldwide through a combination of vector control methods and drug treatment using the antimalarial chloroquine (CQ) [133]. After two decades the focus of these efforts shifted to a less ambitious goal of malaria control due in part to resistance and safety concerns. Recently, resurgence in malaria cases have been observed, primarily in underdeveloped countries and due in large part to the development of resistance mechanisms to nearly all affordable antimalarial drugs and insecticides [134, 135]. Though it has been over 50 years since resistance to CQ was first reported, no affordable replacement has been developed [136, 137]. Perhaps the primary reason that progress has been slow is because malaria is a disease of poverty [138]. There has also been a lack of public funding for tropical diseases, referred to as the ‘90/10 split’, where 90% of the US health-related funding goes to research for only 10% of the worldwide disease burden [139]. Fortunately, interest in malaria control has increased in recent years, including an initiative announced in 2007 by the Bill & Melinda Gates Foundation to eradicate malaria worldwide. Furthermore, through the advent of public-private partnerships (PPP’s), pharmaceutical companies are now collaborating with non-profit organizations and academic institutions to develop new antimalarial drugs [140].

Although malaria has historically been a neglected disease, there is reason to be optimistic. Several significant advances have been made in antimalarial drug discovery, particularly in the past ten years. Perhaps one of the most exciting advancements has been the full genomic sequencing of *P. falciparum* and *P. vivax* which allows for the identification of new druggable targets [141, 142]. Additionally, a more in-depth understanding of parasite biochemistry has allowed the introduction of high-throughput

screening (HTS) techniques for both target-based and phenotypic assays. Successful target-based assays have been developed including those that identify inhibitors of hemozoin formation, aminopeptidase activity, the mitochondrial electron transport chain, and pyrimidine biosynthesis [143-146]. Phenotypic assays have also been developed to identify molecules that are toxic to parasite cultures. Recently, phenotypic screens have been completed by GSK and Novartis against millions of compounds in their respective HTS collections [147, 148]. The results were released to the public and identified thousands of potent chemical starting points for lead compound development. The scientific community now has access to an arsenal of thousands of compounds that are highly toxic to the parasite, but the process of developing these hits into robust lead compounds will be challenging. Identification of the molecular targets responsible for activity could aid in prioritization of these hits, while the lead optimization could be streamlined since the hit-target interactions could be directly studied, saving time, and money. While the experimental determination of targets is possible in some cases, it is an expensive and time-consuming process. With the wealth of target-based HTS data obtained in recent years, there is an opportunity to utilize these results *in silico* as a training set for the development of a QSAR model capable of predicting target-specific activity against phenotypic data, thereby bridging these two screening techniques. However, the validity of a QSAR model is only as valuable as the quality of the training set used to build the model.

One of the most studied *P. falciparum* drug targets is hemozoin, the malaria pigment [149]. During the intraerythrocytic stages of infection, the parasite consumes hemoglobin as a source of nutrition. Upwards of 80% of an erythrocyte's hemoglobin can be consumed by the growing parasite, resulting in the liberation of near molar amounts of toxic free heme in the parasite's acidic digestive food vacuole. To escape free heme toxicity, the parasite has evolved a mechanism by which heme is incorporated into a non-toxic crystalline material known as hemozoin. Inhibition of hemozoin formation by quinoline drugs such as CQ leads to accumulation of near molar concentrations of toxic free heme, resulting in parasite death. Resistance to the quinoline drugs is the result of efflux mechanisms developed by the parasite

rather than changes to the pathway itself. For this reason, hemozoin remains a valid drug target for the development of new antimalarials. Recently, the NP-40 β -hematin (synthetic hemozoin) formation assay was developed to mimic the conditions present within the digestive food vacuole. An HTS pilot screen using the β -hematin formation assay identified 161/38,400 compounds as actives, and over 30% of these were also active against in vitro cultures of *P. falciparum*. The high translation of target-specific actives into in vitro antimalarial compounds is supportive of a high quality screen, and suggests this assay would be useful in obtaining a training set for the development of a robust QSAR model. In this work, we demonstrate an affordable *in silico* screening approach to sort and prioritize compounds according to their likelihood to interact with a specific target, hemozoin, from the TCAMS data set from GSK.

Results

HTS identification of β -hematin inhibitors

(This section was kindly provided by Rebecca D. Sandlin)

The development of quantitative structure activity relationship (QSAR) models has shown practical value for *in silico* screening of potential hit compounds to identify novel chemical entities with a desired biological activity [42, 82, 84] and prioritize focused libraries for acquisition and experimental verification. The previously reported NP-40 β -hematin formation assay represents a high quality assay to serve as the training set for development of the QSAR model. The assay was designed to mimic biological conditions where hemozoin formation occurs in order to enrich the subset of β -hematin inhibitors that translate into in vitro antimalarial compounds [144]. Upon screening 144,330 compounds, 530 β -hematin inhibitors were identified with more potent activity than CQ. Follow-up dose-response curves revealed potencies ranging from 0.5 to 110 μ M against heme-crystallization (see Supplementary Information). Examination of the structures that were identified as having activity against β -hematin formation highlights 24 scaffolds that represent 321 of the 530 total compounds identified in this screen (see Figure 2).

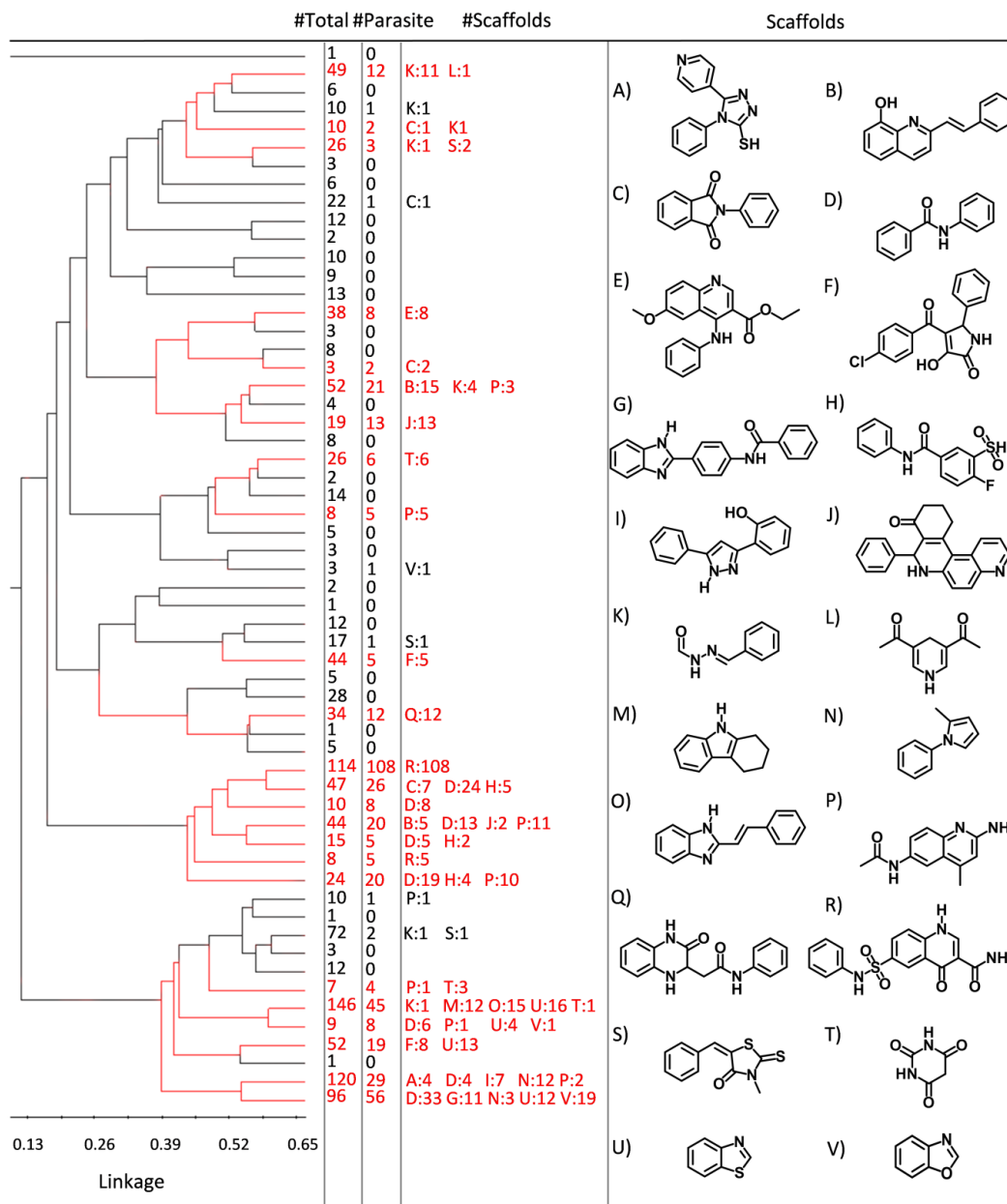


Figure 2: Dendrogram representing the majority of cluster scaffolds of 530 β -hematin inhibitors. Dendrogram branches in red denote scaffolds that kill the parasite and show β -hematin inhibition.

A search of the PubChem database [44] revealed that compounds derived from many of the 24 scaffolds identified in the clustering analysis have previously been reported as in vitro antimalarials. However, no previously reported activity against β -hematin was found for scaffolds 2, 4, 6, 8, 10, 11, 16, 17, 18, 20, 22

and 23. The remaining unassociated 209 compounds encompass a wide range of structural diversity (see Appendix).

Confirmed hits were further examined for *in vitro* antimalarial activity, leading to the identification of 171 compounds with activity against the D6 strain of *P. falciparum*. IC₅₀ values ranged from 0.11 – 17.8 μM. This high translation of target-specific actives (32%) compares favorably to previously reported assays where only 3% of β-hematin inhibitors exhibited *in vitro* antimalarial activity. By utilizing an assay that successfully recapitulates the biological conditions within the digestive food vacuole of the parasite, a higher percentage of compounds retained activity against *in vitro* cultures of *P. falciparum*, suggesting hemozoin formation as the pathway. This assay design also facilitated the discovery of novel and highly diverse scaffolds with potent β-hematin inhibitory activity. Consequently, we predicted that the results of this screening effort would form the basis of an excellent training set for the development of an accurate QSAR model to predict pathway specific activity against hemozoin formation from libraries of phenotypic actives.

QSAR model development

Here, the results of the high throughput screen have been harnessed to provide the knowledge-base for a target specific QSAR model developed with BCL::ChemInfo, an *in silico* ligand based virtual screening suite. A detailed description is provided in the methods section. The predictive power of the QSAR models was assessed by means of quality measures such as precision, enrichment, information gain, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve for an independent dataset. The best performing consensus model was chosen by maximizing its precision in the range critical for *in silico* screening and prioritization of compounds from a library not used for model training. The final consensus model achieved an average precision of 16% for the top 0.1% - 1% of positive predicted compounds, i.e. 16% of the predicted active compounds are true actives. This corresponds to an enrichment of 44 fold in comparison to the initial hit rate (0.37%) when using a cutoff of 1% false positive rate (see Appendix).

QSAR model successfully prioritizes GlaxoSmithKline and Novartis compounds for inhibition of β -hematin pathway

GSK and Novartis screened their small molecule libraries and identified hit compounds acting against in vitro cultures of *P. falciparum*. The targets of these potent in vitro antimalarial compounds are largely unknown and it would be far too expensive to individually identify the targets of each of the hits. In an effort to prioritize compounds most likely to have activity against the hemozoin formation pathway, the QSAR model was applied to virtually screen the active compounds for β -hematin inhibitory activity published by GSK and Novartis. The output of the QSAR model is a predicted IC₅₀ (P-IC₅₀) for β -hematin inhibition. The compounds were grouped into seven 'bins' based on the potency of the predicted IC₅₀ values (0-70 μ M, 71 - 140 μ M, etc). The GSK library contained a total of 13,229 compounds of which 249 compounds (below 70 μ M) were predicted to express inhibitory effects *in silico*. The Novartis library contained 5,697 compounds and the computational model predicted 37 compounds as inhibitors with an IC₅₀ of below 70 μ M. These results yield a predicted *in silico* hit rate of 1.9 % and 0.65% for the GSK and Novartis compounds, respectively.

Experimental analysis of enrichment by predicted activity binning and structural similarity

(This section was kindly provided by Rebecca D. Sandlin)

The QSAR model was used to prioritize compounds based on predicted β -hematin inhibitory activity. If the QSAR model is indeed accurate, experimental analysis should confirm a significant number of the compounds in bin 1 to be active while bin 9 compounds should be inactive, such that the highest degree of enrichment should be observed in bin 1. To fully validate the accuracy of the QSAR model to prioritize inhibitors of β -hematin formation, the entire library of 13,229 GSK in vitro antimalarial compounds was experimentally analyzed. Compounds were first pre-screened in duplicate using the NP-40 assay at a concentration of 22 μ M. Using an activity threshold of >80% β -hematin inhibition, 397 hits were identified (3% hit rate) from the 13,229 compound library. These hits were then compared to the list of predicted hits identified *in silico* (P-IC₅₀ \leq 70 μ M) to establish the accuracy of the developed QSAR

model. Of the 249 in vitro antimalarials predicted to inhibit β -hematin formation, 40 were confirmed active (16% hit rate), an enrichment factor of 44 compared to the initial training set used to develop the QSAR model. Bins 2 and 3 also showed increased degrees of enrichment where 10% and 11% of the compounds tested were active, respectively. In bins 5-9, the percentage of actives decreases as a function of increasing bin number, such that the largest ratio of active:inactive is observed in bin 1, the bin predicted to contain the highest percentage of actives according to the QSAR model.

It is also worth noting that some degree of enrichment is observed in bins 5-9. This is almost certainly due to the fact that the GSK and Novartis compounds are biased as they are known to inhibit some pathway within the parasite. As such, a higher hit-rate would be expected than that observed when screening a non-biased library, leading to some degree of enrichment (the overall hit-rate of the GSK library was 2%, higher than the 0.37% hit-rate in the unbiased HTS library used as the training set).

Comparison of the QSAR model to naïve substructure search

In addition to the development of the QSAR model, a naïve substructure search (NSS) was evaluated to assess the structural similarity of all 13,229 GSK compounds compared to the known 530 active compounds used in QSAR model training (see Figure 3) and to determine whether a less sophisticated model prediction approach could identify β -hematin inhibitors. 89% of all GSK compounds and 91% of all experimentally validated compounds have a similarity <50% similarity to the known actives. Considering compounds with a similarity cutoff of >80% to the known training actives the naïve substructure search identified <1% of the total number of compounds, containing <2% of all experimentally validated active compounds highlighted in the red partition. In comparison, the bin 1 compounds identified by the QSAR model (green partition) identified <2% of the total number of compounds but contained 10% of the total number of actives. Bin 1 compounds also exhibited a significantly higher hit-rate of 16%, an enrichment factor of 44 compared to the initial training set, demonstrating that the QSAR model is significantly more powerful at identifying populations of compounds with enriched activity against the selected target, β -hematin.

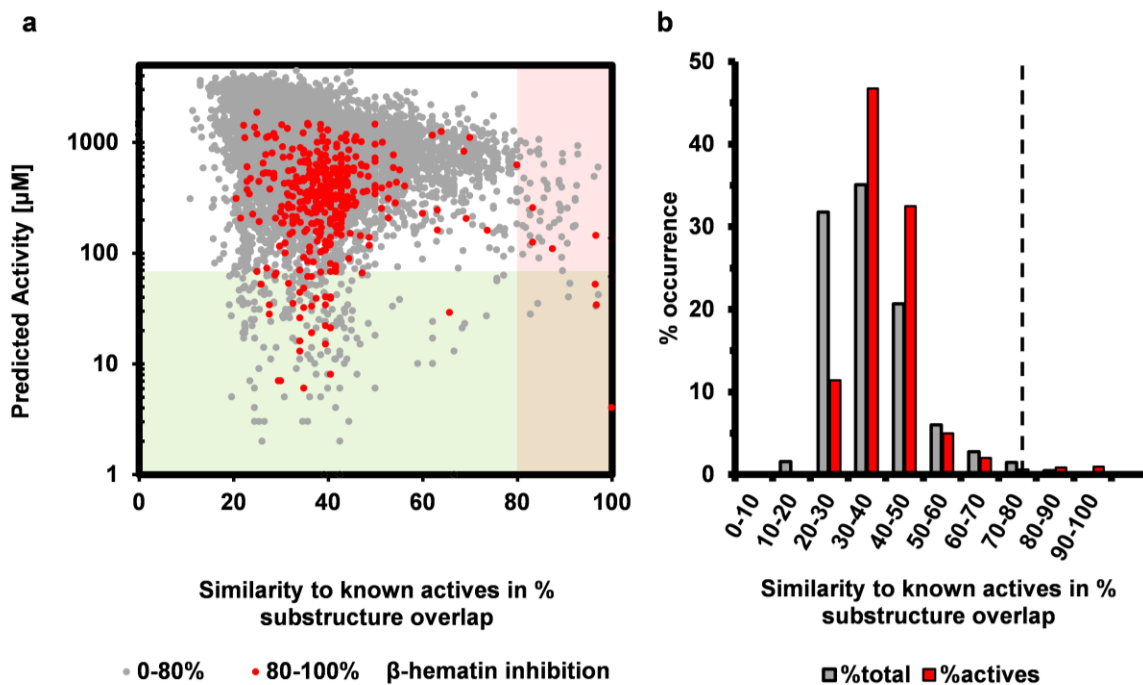


Figure 3: QSAR model predictions compared to naïve substructure search

- Visualization of the predicted activity of GSK compounds by the QSAR model in relation to the similarity of the 530 known active compounds by % substructure overlap. Compounds represented in grey exhibited a β -hematin inhibition of less than 80% and are classified as inactive. A red color coding indicates an experimentally validated active compound. The green partition corresponds to bin 1 with concentration threshold of $<70\mu\text{M}$. The red partition denotes all compounds that express a similarity of $>80\%$ by substructure overlap. The area where the green and red partitions overlap contains compounds that were identified by both prediction approaches.
- Distribution of GSK compounds with respect to similarity towards known actives used in QSAR model training in % substructure overlap. The dotted line corresponds to the beginning of the red partition in A) distinguishing between compounds having similarity of less or greater than 80% by substructure overlap.

Target validation in *P. falciparum*

(This section was kindly provided by Rebecca D. Sandlin)

As evident in the screening results, many of the in vitro antimalarial compounds identified by GSK were also active in the β -hematin formation assay, suggesting hemozoin as the target. In order to validate hemozoin as the molecular target responsible for parasite death, the heme speciation assay developed by Combrinck et al. was utilized to quantitate hemozoin formation [150]. In this assay, tightly synchronized ring-stage parasite cultures are treated with the indicated β -hematin inhibitor and incubated for 32 hours, allowing the rings to mature into schizonts where hemozoin formation is expected to be maximal. Parasitic digestive food vacuoles are subsequently isolated from culture, facilitating the quantification of free heme and hemozoin concentrations. If the β -hematin inhibitor indeed targets the hemozoin formation pathway, then the ratio of free heme:hemozoin should increase as a function of increasing drug concentration and consequent parasite death. Figure 4 shows the treatment of parasite cultures with TCMDC-123692, confirming that upon treatment, hemozoin is inhibited, leading to the accumulation of free heme and consequent parasite death. In total, 17 β -hematin inhibiting in vitro antimalarials were tested in this assay, 15 of which were confirmed to inhibit hemozoin formation in *P. falciparum*.

Methods

β -hematin formation HTS assay

(This section was kindly provided by Rebecca D. Sandlin)

The β -hematin formation assay has been described previously for use in HTS [144]. A bulk liquid dispenser was used to add water, NP-40 (30.6 μ M), acetone and heme (100 μ M) suspended in a pH 4.9 acetate buffer (1 mM) to 384-well microtiter plates. The plates were incubated for six hours in a shaking water bath at 37°C. The pyridine-ferrochrome method of colorimetric quantification was used to analyze the inhibitory activity of compounds by adding a solution containing 50% pyridine, 20% acetone, and water solution buffered with a 200 mM pH 7.4 HEPES buffer such that the final concentration in each well was 5% pyridine (vol/vol) [151]. From this method, hits were then identified as compounds exhibiting greater than 80% inhibitory activity. Follow-up dose-response curves were generated for in a concentration range of 0.1-110 μ M.

QSAR model training

Quantitative structure activity relationships seek to correlate the often complex non-linear relations between chemical structure and biological activity for a specific protein target [82, 84]. Modern QSAR techniques employ advanced 2D molecular fingerprints and 3D molecular descriptors coupled with machine learning [152, 153]. The descriptors employed in this study (scalar, 2D/3D auto-correlation, radial distribution functions) are fragment-independent and translation/rotation invariant. Thus, the resultant fingerprints are constant in length and independent from molecule orientation in space. A total of 1,284 numerical descriptors in 60 categories were implemented in this study (see Supplementary Information). 3D conformations were calculated for all molecules with CORINA [12]. A set of five machine learning algorithms was implemented in the BCL::Cheminfo, a virtual screening suite for ligand-based computer aided drug discovery. This study employed machine learning techniques such as Artificial Neural Networks [22, 154, 155], Support Vector Machines [21], kappa – Nearest Neighbors [156, 157], Kohonen Networks [98], and Decision Trees [158]. Machine learning models can adapt to complex interrelations and are capable of detecting even small signals at high noise levels. Hence, each of these methods is applied when no simple mathematical model can be assumed, many influencing factors interact, and the experimental uncertainty is high. The training data set had employed 80% of the data points used in the training process. The number of training iterations was limited through early termination to counter “overfitting” of the machine learning model to the training data. A monitoring data set consisting of 10% of the data points was used to optimize all training parameters of the machine learning methods and to invoke early termination. The final 10% of the data points are set aside as an independent data set. It is not used in the training process, but used to evaluate the final model. There was no overlap of compounds between training, monitoring, and independent data sets. An optimized descriptor set was determined for every combination of machine learning technique by means of information gain of each numerical descriptor element. Systematically, each data partition is chosen once as an independent dataset and cross-validated through machine learning model training [159]. The overall

accuracy of all cross-validated QSAR models was evaluated by calculating a jury consensus of all predicted P-IC50 values and each independent data set iteration.

Heme speciation assay

(This section was kindly provided by Rebecca D. Sandlin)

The heme speciation assay was conducted using the method previously described. Briefly, a highly synchronized culture of *Plasmodium falciparum* (3D7) in the early ring stage was evenly divided into culture flasks and titrated with indicated β -hematin inhibitor. The cultures were allowed to incubate (37°C, 5% O₂, 5% CO₂, 90% N₂) for 32 hours to mature into the late trophozoite stage of the parasite lifecycle where hemozoin formation is maximal. Following incubation, the trophozoites were isolated using saponin lysis (0.05%) and collected for further quantification of the three forms of heme: hemoglobin, free heme, and hemozoin. To the trophozoite pellet, HEPES buffer (0.02M, pH 7.5) and sodium dodecyl sulfate (4%) was added, followed by sonication and centrifugation. The supernatant was collected as the hemoglobin fraction, while pyridine (5%) was added to the pellet to solubilize the free heme. The remaining fraction consisted of hemozoin, which was then solubilized through the addition of sodium hydroxide (0.3M). For each fraction, the UV-visible absorption spectrum was collected between 300 and 800nm with the maximum peak maximum at 405nm used to calculate the percentages of the heme species.

Discussion

The GSK and Novartis screening efforts resulted in the identification of thousands of chemical starting points for lead identification, but the molecular targets of these active compounds are unknown. Leveraging this rich data source for the development of lead compounds hinges on identification of pathway and target, a time-consuming and resource-intensive process. Hemozoin, a validated target specific to the parasite, has been studied in great detail and serves as an important pathway for future antimalarial development. Inevitably, some subset of the in vitro antimalarials identified by GSK and Novartis were likely to act upon the hemozoin pathway. To avoid testing all 20,000 active compounds, the challenge is to prioritize those that are most likely to exhibit activity against a particular target, in this instance,

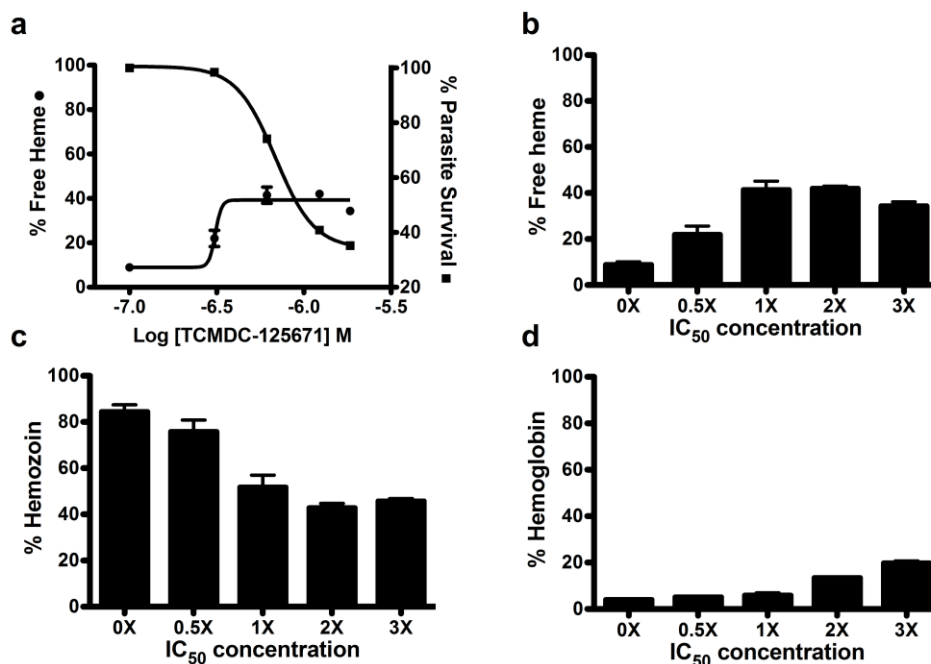


Figure 4: Validation of β -hematin inhibitors in *Plasmodium falciparum*.

(a) Parasite survival curve (squares, right axis) determined using the malaria SYBR green I fluorescence based assay and the percentage of free heme (circles, left axis) as a function of compound concentration. Fraction of the (b) free heme, (c) hemozoin, and (d) hemoglobin as a percentage of the total heme species present following treatment with TCMDC-125671.

hemozoin formation. This has been addressed here using a QSAR model to virtually screen the 20,000 compounds to prioritize those that are most likely to inhibit β -hematin formation. The model was subsequently validated through the experimental analysis of over 13,000 of the GSK compounds to determine the feasibility of screening *in silico* libraries in the future.

The training set was provided by an HTS campaign that screened 144,330 compounds to identify inhibitors of β -hematin formation (0.37% hit-rate). This method of developing a QSAR model takes advantage of the chemical space provided by both active and inactive compounds. The resulting model was used to virtually screen the GSK and Novartis hits and prioritize the compounds based on P-IC50 values of ≤ 70 μ M. This approach prioritized a manageable subset of 249 GSK and 37 Novartis compounds with predicted activity against β -hematin formation. To fully validate the accuracy of the QSAR model, the entire GSK *in vitro* antimalarial library was experimentally tested for activity against β -hematin formation. Of the 249 compounds in Bin 1 (those predicted to have the most potent activity against β -hematin formation), 40 compounds were experimentally confirmed active, an enrichment of 44 fold. Compounds with a low P-IC50 value are predicted with a high confidence yielding a decrease of enrichment with increasing bin number. Therefore, this method confidently filters for inhibitors of β -hematin formation.

A comparison between the QSAR model and a NSS was conducted to determine whether a less sophisticated prediction approach could be equally successful for predicting GSK compounds that inhibit β -hematin formation. Using the NSS, <1% of the GSK compounds exhibited >80% similarity to the 530 known actives that were used in the training set. Since only 9 of these compounds were true actives, this approach is limited in its ability to predict inhibitors of β -hematin formation. In contrast, the QSAR model identified 40 compounds from Bin 1 as actives. Furthermore, the compounds predicted by the QSAR model explore a larger chemical space and are more suitable to identify antimalarials comprising new chemotypes.

The QSAR approach developed here is not specific to the malaria parasite, and can be used for target-elucidation in other diseases as well. This is particularly applicable for use in neglected diseases where limited funding restricts the expensive and time-consuming process of target-elucidation. A further application of this model would be to guide the selection of second generation compounds in the process of lead development. By creating a large list of possible second generation compounds, the model can prioritize a subset of compounds predicted to have more potent activity against the target. This approach could greatly accelerate the identification of drug targets and optimization of leads for tropical neglected diseases where research and development funds are severely limited.

4. NOVEL ALLOSTERIC MODULATORS FOR MGLU₅ RELATED TO CPPHA BINDING

Introduction

Metabotropic glutamate receptors (mGlu) are part of the G protein-coupled receptor (GPCR) family also known as seven-transmembrane (7TM) domain receptors. As the primary excitatory neurotransmitter in the mammalian nervous system glutamate acts as an endogenous ligand to mGlu receptors binding orthosterically at a large extracellular globular domain. mGlu activation is tied to a downstream effector mechanism through guanine nucleotide binding proteins. The mGlu family consists of eight subtypes classified into three major groups [55, 160]. Group I includes mGlu₁ and mGlu₅, which couple primarily to G_{q/11} and mediate IP₃/Ca²⁺ signal transduction while increasing phosphoinositide hydrolysis in expression systems [161] whereas group II (mGlu 2 and 3) and group III (mGlu 4, 6, 7, and 8) mGlu receptors couple primarily to inhibition of adenylyl cyclase when expressed in cell lines [162]. Through the diverse physiological roles, heterogeneous distribution, and wide diversity of mGlu receptor subtypes, the possibility arises for new therapeutic agents. These agents are selectively interacting with mGlu receptors specific to a single or limited number of CNS functions. Such therapeutics could have incisive importance on the development of novel treatment strategies for a range of neurological disorders [57, 163, 164].

Modulators of mGlu₅ have promising potential for treatment of schizophrenia

Modulators of mGlu₅ are of particular interest. An increasing body of studies provides evidence for modulators of mGlu₅ to play key roles in potential novel treatment strategies for schizophrenia, Alzheimer's and other diseases linked to cognitive impairment [165, 166]. Current mGlu₅ positive allosteric modulators (PAM) have been developed based of multiple scaffolds [60, 167-175]. Well-characterized structural classes of mGlu₅ PAMs include 1-(3-fluorophenyl)-N-((3-fluorophenyl)-methylideneamino)-methanimine (DFB), 3-cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl)-benzamide

(CDPPB), and N-{4-chloro-2-[42]phenyl}-2-hydroxybenzamide (CPPHA). Challenges with mGlu5 PAMs in medicinal chemistry are characterized by slight structural changes inducing a large spectrum of pharmacological responses (DFB series), 'flat' structure–activity relationship (SAR) (CPPHA series) resulting in a low hit rate in assay experiments, or minimal SAR improvements beyond the initial starting structure (CDPPB series) [162]. Overcoming these difficulties led to evidence supporting these positive modulators to allosterically potentiate mGlu5-mediated electrophysiological responses in the CNS [167, 169, 176, 177]. Furthermore, antipsychotic-like effects were confirmed by in vivo studies based on animal models [60, 178].

More recently, a distinct mGlu5 PAM chemotype was presented by Addex pharmaceuticals (ADX47273) [179]. The compound showed in vivo efficacy in preclinical behavioral models comparable to other known antipsychotics [180]. In general, these mGlu5 PAMs improved cognitive function in animals in which object recognition was impaired [180] and demonstrated improved behavioral flexibility [181]. These exciting discoveries provide a strong emphasis for mGlu5 PAMs having potential application as novel antipsychotic agents and cognition-enhancing therapeutics.

Allosteric modulation of mGlu₅ through multiple distinct binding sites

The primary binding site of these PAMs correlates with the same site as for the negative allosteric modulator (NAM) 2-methyl-6-(phenylethynyl)-pyridine (MPEP) located in TM 3,6, and 7 [165]. Evidence for a possible distinct allosteric binding site was provided by O'Brien et al. [167]. In a later study, Chen et al [168] characterized a novel mGlu₅ PAM labeled 4-nitro-N-(1,3-diphenyl-1H-pyrazol-5-yl)-benzamide (VU-29) which was shown to act at an overlapping site to the binding site of MPEP. The same study points out that the structurally distinct PAM CPPHA potentiates the concentration response of mGlu₅ by an alternative mechanism to the one of VU-29. Additionally, a neutral ligand at the MPEP allosteric site termed 5-methyl-2-(phenylethynyl)-pyridine (5MPEP) is blocking VU-29- and CPPHA-induced potentiation of mGlu₅ responses. Remarkably, an increase in 5MPEP concentration invokes a parallel rightward shift in the concentration-response curve of VU-29. In contrast, CPPHA potentiation is

inhibited by 5MPEP noncompetitively [169, 177]. Furthermore, this evidence of an alternative binding site is supported by mutagenesis experiments where a mutation (A809V) decreases binding of ligands to the MPEP site. The mutation abolishes binding of VU-29 but does not influence the response to CPPHA. Likewise, a second mutation (F585I) eliminates the response to binding CPPHA but does not change binding of VU-29. Both results suggest that CPPHA does not bind at the MPEP site but rather acts at a distinct novel allosteric site on mGlu₅ as a positive potentiator to the receptor.

A later study by Noetzel et al. [182] revealed an analog to the CPPHA series that was investigated and labeled NCFP (N-(4-chloro-2-((4-fluoro-1,3-dioxoisindolin-2-yl)methyl)phenyl)picolinamide). NCFP binds to the previously mentioned CPPHA site on mGlu₅ and exhibits stronger mGlu₅ subtype selectivity in comparison to CPPHA increasing the effectiveness for studies of mGlu₅ influence in the CNS. Furthermore, an analysis to investigate the effects of agonists and antagonists on the cellular response regarding mGlu₅ revealed that NCFP will induce a parallel rightward shift in concentration–response curves independent of the presence of the neural ligand 5MPEP.

Drug discovery guided by *in silico* virtual screening can prioritize modulation-specific mGlu potentiators

Previous quantitative structure activity relations (QSAR) studies were successful in identifying lead scaffolds serving as starting points for successful probe development campaigns. A research study by Mueller et al. [43] identified novel mGlu₅ PAMs based on a high-throughput screen (HTS) of a diverse library of 144,475 substances which revealed 1,382 compounds as PAMs. A subsequently trained QSAR model with a theoretical enrichment ratio of up to 38 for an independent data was applied to screen a database of approximately 450,000 commercially available drug-like compounds. A set of 824 compounds was acquired for testing based on the highest predicted potency values. Experimental validation confirmed 28.2% (232/824) of these compounds with various activities at mGlu₅. These results represent an enrichment factor of 23 for pure potentiation of the mGlu₅ glutamate response and 30 for

overall mGlu₅ modulation activity when compared with those of the original mGlu₅ experimental screening data (0.94% hit rate).

A subsequent study [183] applied the same approach to investigate the activation of metabotropic glutamate receptor subtype 4 (mGlu₄) shown to be efficacious in rodent models of Parkinson's disease. A high throughput screen identified 434 positive allosteric modulators of mGlu₄ out of a set of approximately 155,000 compounds. A QSAR model with a theoretical enrichment of 15 fold when selecting the top 2 % compounds of an independent test dataset. The same external commercial database of approximately 450,000 drug-like was screened as in the previous study. 1,100 predicted active small molecules were tested experimentally using two distinct assays of mGlu₄ activity. This experiment yielded 67 positive allosteric modulators of metabotropic glutamate receptor subtype 4 that confirmed in both experimental systems. Compared to the initial hit rate of 0.3% in the primary screen, the enrichment resulted in a 22 fold increase.

Noeske et al. [184] applied a self-organizing map to discern antagonists for mGlu₁ and mGlu₅. This approach employed topological pharmacophore descriptors which were applied to a compiled library of 338 compounds of non-competitive antagonists and an external library of 5,376 molecules of known drugs and lead candidates for different drug targets. The resulting self-organizing maps were able to differentiate between separate localized distributions for the two mGlu receptor targets.

Significance

Preclinical studies have implied the role of mGlu receptors as possible candidates for drug targets involved in treatment of a range of CNS disorders. Research efforts are focused around the development of allosteric modulators due to receptor subtype selectivity and activity reliance upon receptor activation. However, the current understanding of mechanisms that involve mGlu receptors is limited when mediated *in vivo* effects are considered. In addition, awareness is increasing that distinct allosteric modulators can induce differential effects on mGlu₅-mediated psychological responses in the CNS, referred to as

functional selectivity or stimulus bias [185-187]. Albeit these challenges, promising evidence has demonstrated the existence of a novel mGlu₅ binding site distinct from the MPEP binding site. The identification of novel allosteric potentiators binding to this alternative site will help elucidate binding interactions and bias effects of ligands binding at one of these two sites. Computational techniques like virtual screening are imperative for prioritization and identify of novel selective potentiators. This study identified 63 compounds through virtual screening and rigorous filtering which were purchased and experimentally validated. Approximately 10% (6/63 compounds) were confirmed to be specific allosteric mGlu₅ PAMs active at the CPPHA-site. This represents an virtual screening hit rate increased by two orders of magnitude compared to an initial experimental hit rate of 0.09% (130 CPPHA like actives/(1382 PAMs + 144,475 inactives)). These identified molecules have the potential to serve as lead compounds in future drug discovery campaigns and will subsequently help to establish specifically tailored therapeutics with reduced side effects.

Results

With the identification of a putative allosteric mGlu₅ binding site different to the MPEP binding site a series of CPPHA analogs was screen and accessed for mGlu₅ PAM activity. A total of 275 distinct compounds were screened with 130 compounds tested to be active (see Table 4).

Table 4: Listing with all available datasets used for QSAR modeling and virtual screening study.

Type	#actives	#inactive
mGlu ₅ PAM	1,382	144,475
mGlu ₅ NAM	343	152,298
mGlu ₅ PAM CPPHA	130	145

This data set is comprised of CPPHA-like compounds. Structural similarity between the involved different scaffolds is shown in Figure 5. The dendrogram shows structural similarity of each scaffold

cluster with a representative structure. Additionally, for each cluster the composition of active and inactive compounds is indicated relative to the overall total number of compounds.

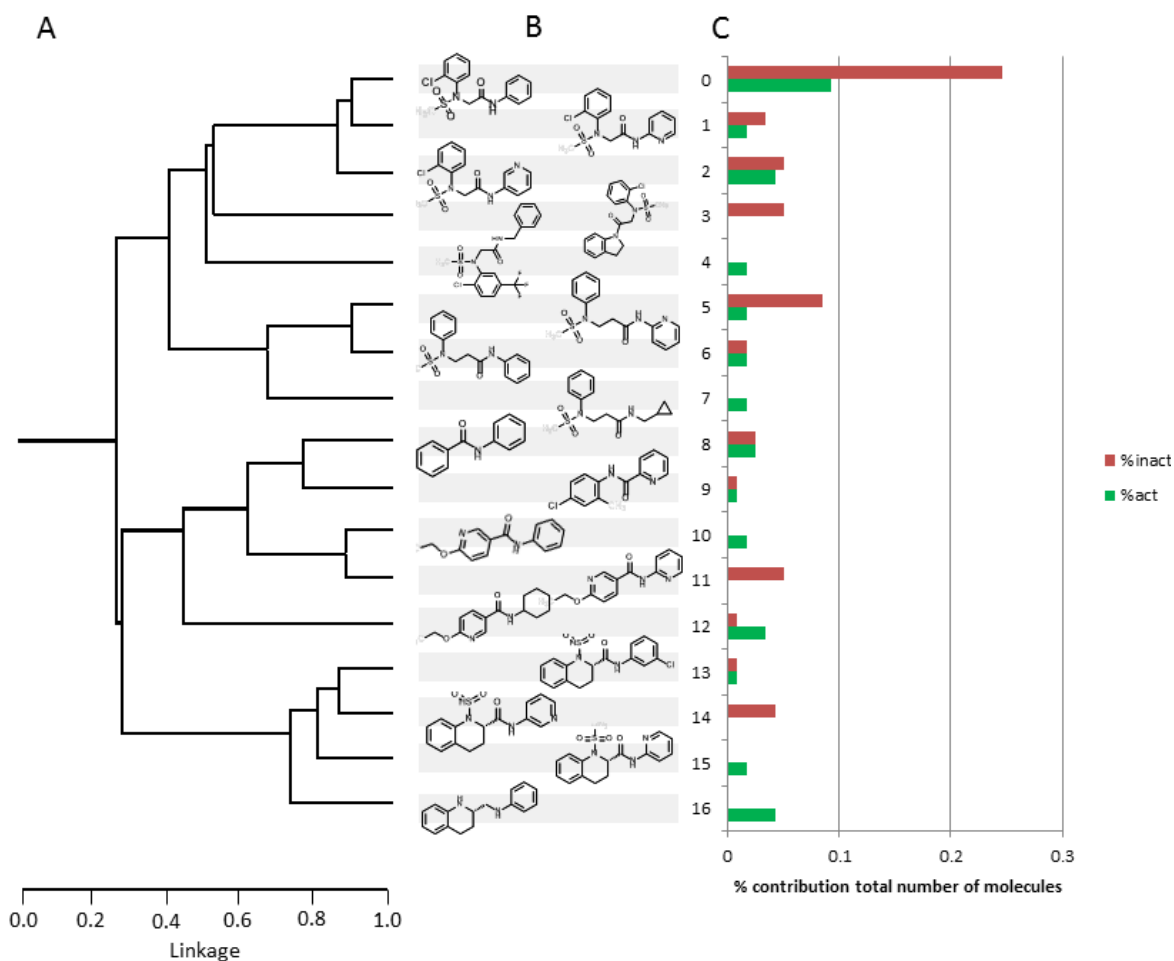


Figure 5: Clustering of compounds 130 actives and 145 inactives in the CPPHA series.

A) shows the structural similarity of 17 identified clusters. B) highlights the scaffold representation of each cluster and C) shows the cluster composition in respect to actives (green) and inactives (red).

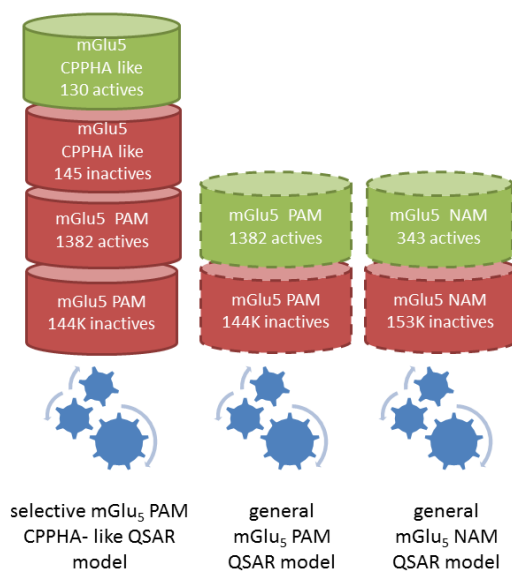


Figure 6: Schematics of data set composition for each of the three QSAR models hit compounds.

Development of QSAR model to prioritize specific mGlu₅ PAMs for alternative site distinct from MPEP binding site

Three cross-validated QSAR models were trained based on the aforementioned datasets. In previous studies we developed two specific QSAR models to predict allosteric mGlu₅ PAMs [43] and NAMs [183] from a HTS campaign screening for mGlu₅ modulators. The same datasets were used, as reported in the respective publications, to re-train two of those QSAR models with our current advancements version of the BCL. In this study, a third QSAR model was developed specifically to identify compounds that bind to the putative CPPHA site. In a previous study [43] a set of 1,382 PAMs was identified to activate mGlu₅ allosterically. This model did not distinguish between selective binders to the MPEP site or CPPHA site. Here, these 1,382 PAMs are classified as inactives because of their non-CPPHA like character. There is no molecule overlap between the set of 130 actives (CPPHA-like) and the set of 1,382 general mGlu₅ PAMs. The dataset for training was comprised of the 130 active compounds (CPPHA-like) categorized as active. All compounds in the remaining dataset were classified as inactive consisting of mGlu₅ PAM actives (1,382 compounds), mGlu₅ PAM inactive (144,475 compounds), and mGlu₅ PAM

CPPHA-like inactives (145 compounds). The mGlu₅ PAM actives (1,382 compounds) were categorized as inactive in this scenario to identify CPPHA-site binders and not binders that interact with the MPEP binding site. The data set configuration for all involved QSAR models is presented in Figure 6. Given this data set arrangement it was possible to develop QSAR models that distinguish between mGlu₅ PAMs, NAMs, and PAMs that bind the alternative CPPHA binding site distinct from the MPEP binding site. The cross-validated QSAR models were evaluated by means of the receiver operating characteristics (ROC)

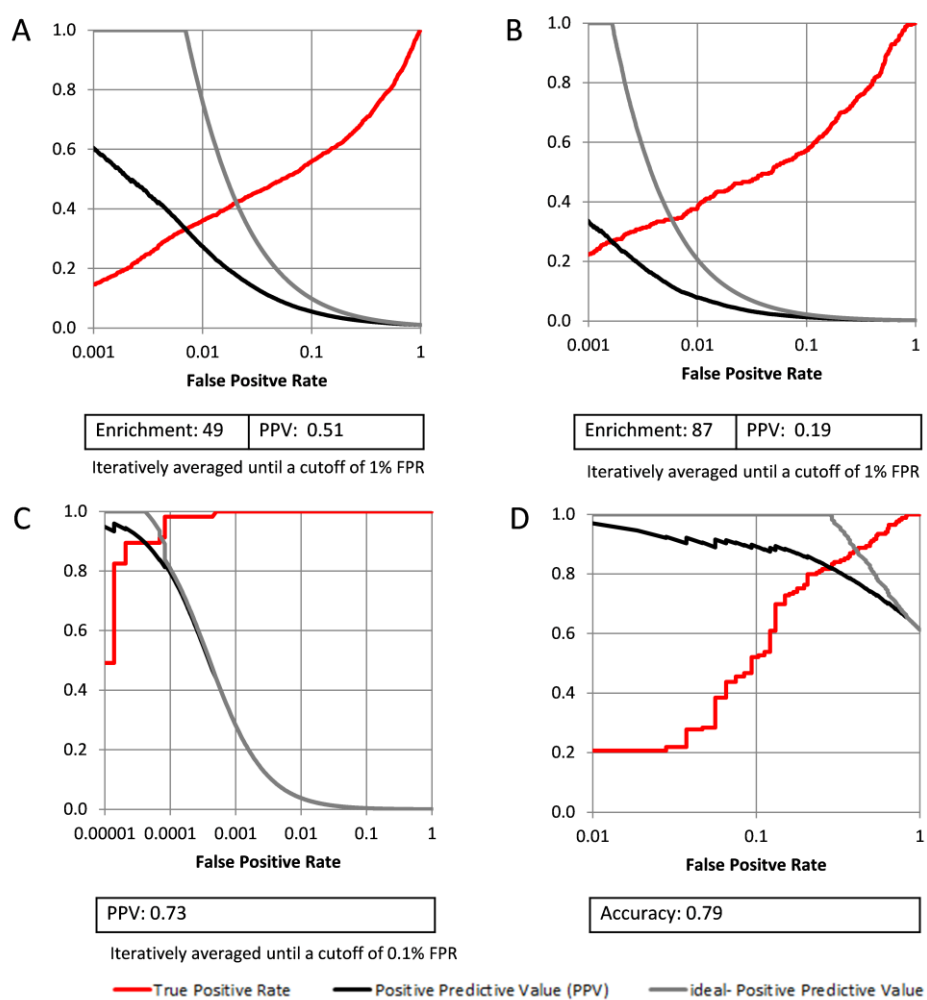


Figure 7: Listing of ROC curve quality measures for each cross-validated model on a log scale.

A) general mGlu₅ PAM model. B) general mGlu₅ NAM model. C) ROC curve for the specific mGlu₅ PAM model predicting CPPHA-site compounds. D) ROC curve when only the 130 actives / 145 inactives portion of the independent data set is considered.

curve and enrichment (ENR). The enrichment and positive predictive value (PPV) results were calculated by iteratively averaging the quality measure up until a specific false positive rate (FPR) cutoff for each of the three QSAR models. The results are shown as ROC curves in Figure 7. The selective mGlu₅ PAM model for CPPHA-site binders achieved a PPV of 73% with a prediction accuracy of 79% when considering only the CPPHA-like actives and inactives in the independent data set.

Virtual screening predicts site specific mGlu₅ PAMs among commercially available compounds

The compound library *eMolecules* [188] was chosen for virtually screening with the aforementioned QSAR models. It contains over 4 million compounds from several large vendor libraries. To query for selective hits of the mGlu₅ PAM CPPHA binding site, the three QSAR models were applied to simultaneously virtual screen for activity of each respective mGlu₅ target. An mGlu₅ PAM CPPHA-site hit was defined by molecules classified as inactive by the general mGlu₅ PAM model, inactive by the general mGlu₅ NAM model, and active by the selective mGlu₅ CPPHA-site PAM model. Through this query scheme, 4,719 compounds were predicted to be hits.

Medicinal chemistry filters and clustering analysis narrow down selective hit compounds. All 4,719 hits were further assessed by the Rapid Elimination Of Swill (REOS) filter [189] and Pan Assay Interference Compounds (PAINS) filter [190] to remove undesired non-drug-like compounds and often seen false positive molecules. The resulting set of compounds contained 353 molecules which were further filtered by the number of occurring halogen atoms. The final set contained 138 molecules, 109 with one halogen atom and 29 with no halogen atoms. A subsequent hierarchical cluster analysis narrowed this set of molecules further down to 103 candidates, each representing a cluster of structurally similar molecules where the candidate compound with the highest predicted activity for mGlu₅ PAM CPPHA-site binding was chosen. Taking pricing and availability into consideration, a subset of 63 of 103 identified candidate molecules was chosen for acquisition. To further investigate the chemical diversity among the chosen 63 candidate compounds to the known 1,382 PAMs, an analysis of the similarity between every compound pair was conducted (see Figure 8). The distribution of the highest similarity of each compound pair

clearly indicates a scaffold similarity peak at approximately 40%. The majority of compound pairs exhibit a scaffold similarity below 50% which is indicative of scaffold diversity among the chosen compounds.

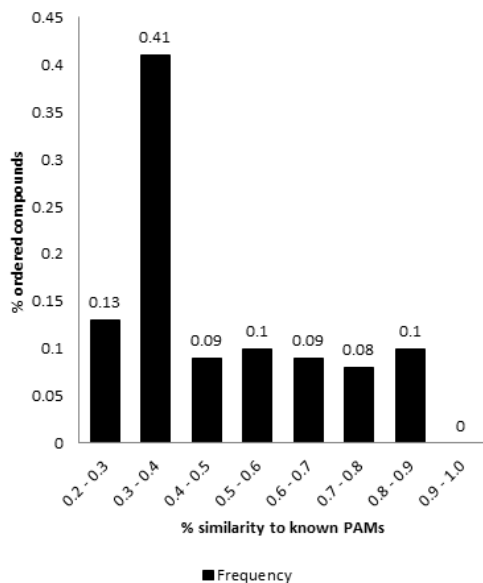
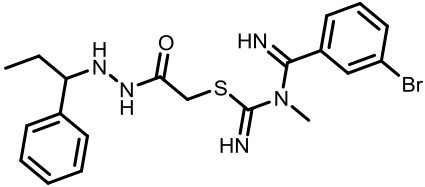
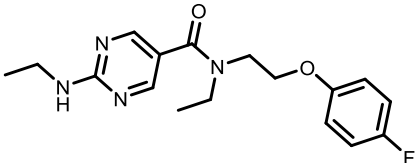
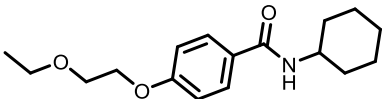
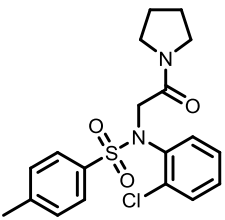
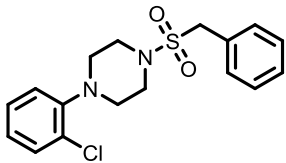
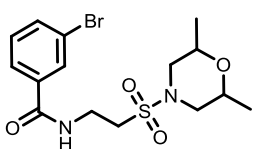
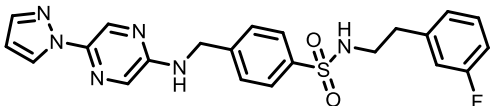


Figure 8: Distribution plot shows the similarity of ordered compounds compared to known PAMs.

Experimental validation of identified compounds through vHTS

All 63 purchased compounds were evaluated by means of calcium binding. An initial testing applied a 10-point concentration response curve (CRC) format. A total of 7 compounds showed an increased response of mGlu₅ in rat brain tissue. The selected compounds were part of follow-up experiments in replicate assays. A second pass of evaluation involved re-plating and re-testing again in 10-point CRC format throughout three independent assays. Results were averaged and seven compounds showed significant activity upon repeat. Table 5 shows a list of all seven identified modulator candidates. PAM activity of the top three compounds reached a glutamate response of 40 to 50% to the maximal signal with potencies above 10 μ M shown in Figure 9.

Table 5: Seven confirmed mGlu5 modulators predicted to interact with CPPHA binding related site.

Structure	VU Identifier	Category	Potency
	VU0603815	Potentiator	6.30E ⁻⁰⁷
	VU0603849	Potentiator	1.44E ⁻⁰⁶
	VU0603805	Potentiator	1.51E ⁻⁰⁶
	VU0603830	Potentiator	1.64E ⁻⁰⁶
	VU0603838	Potentiator	1.77E ⁻⁰⁶
	VU0603841	Potentiator	1.00E ⁻⁰⁵
	VU0603820	Weak Antagonist	1.00E ⁻⁰⁵

Seven compounds were tested in triplicate using membranes harvested from the same mGlu₅ cell line used for the aforementioned calcium assay. A concentration range up to 300 μM was tested (see Figure 10). It was confirmed that five compounds (VU0603830, VU0603841, VU0603838, VU0603815, and VU0603820) showed no radio ligand displacement up to the highest concentration tested, indicating they do not bind to the MPEP site. Compounds VU0603805 and VU0603849 exhibited weak binding at concentrations of 100 and 300 μM. PAM activity at 10 and 30 μM was observed indicating a disconnect between potentiator activity and binding, suggesting an alternate or overlapping binding site.

Methods

Machine learning applied to develop QSAR

A large fraction of proteins targeted by therapeutics have no suitable structural model for structure-based virtual screening. Quantitative structure activity relationships (QSAR) are an area of computational research that builds *in silico* models to predict the biological activity for small organic molecules to a protein target [153]. The models seek to quantitatively correlate complex non-linear relations between

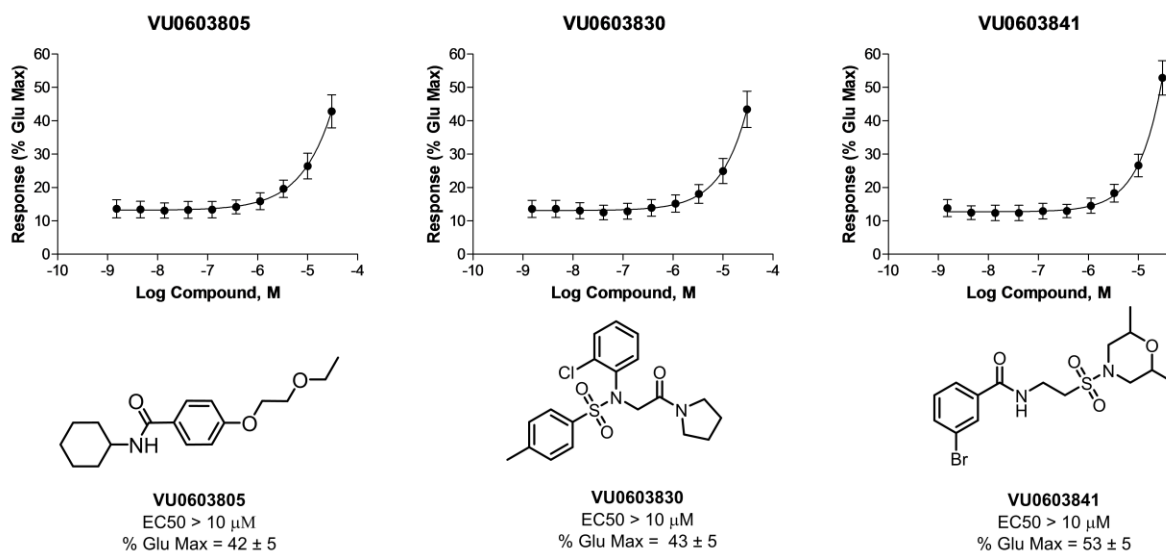


Figure 9: Experimental validation of the three most active compounds by calcium binding.

chemical and physical properties of a molecule with its biological response for a specific biological target. The application of QSAR models plays an important role in a variety of fields like drug discovery [191] and toxicity predictions [10]. Knowledge acquired from compounds already biologically tested directs the design of future molecules.

Molecular descriptors play a fundamental role in encoding chemical information contained in a molecule [79]. Through mathematical procedures the chemical information contained within a symbolic representation of a molecule is transferred into a numerical code. A list of all applied molecular descriptors is available in supplemental materials. Each encoding function is independent of translation or rotation of the molecule. They can be readily applied to conformational ensembles and yield a numerical description of constant length independent from size and composition of the substance.

Machine learning approaches have shown to have exciting potential in estimating biological target data.

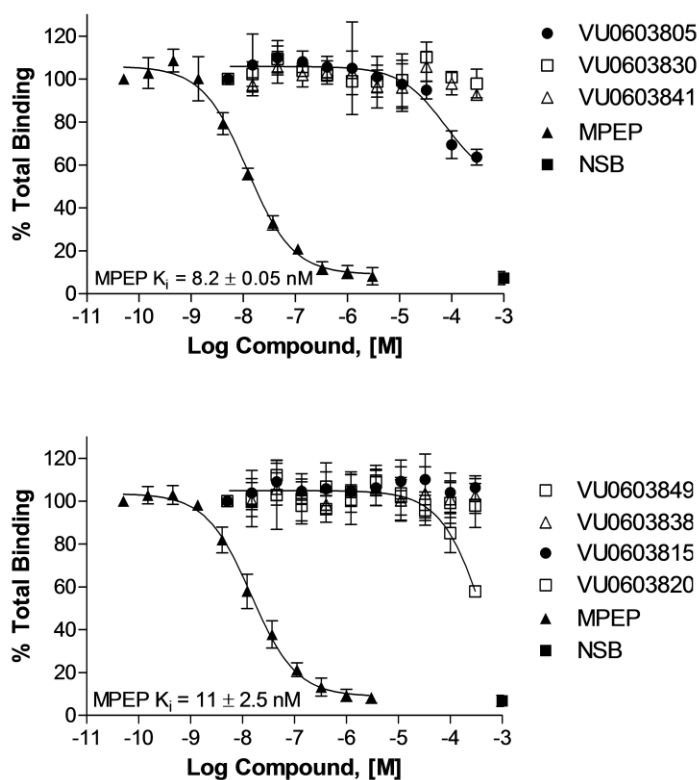


Figure 10: Confirmation of CPPHA site specific binding through radio ligand binding assay.

In recent years the potential of approaches such as Artificial Neural Networks (ANN) for establishing highly non-linear relations has become apparent [22]. The algorithms learn to recognize complex patterns and make intelligent decisions based on an established compound library. Imposing such acquired sets of patterns obtained by a learning process, the algorithms are able to recognize not yet tested molecules and categorize them towards a given outcome. ANNs consist of an interconnected group of artificial neurons and process information adaptively using restricted set of transformations. ANNs have been used for several years in chemistry and biochemistry to describe QSARs [17]. Their critical advantage compared to linear methods such as multiple linear regression lies in the flexibility of the models. ANN models can adapt to complex interrelations and are capable of detecting even small signals at high noise levels.

Clustering of identified hits from virtual screening

To access the similarity of predicted compounds a hierarchical clustering approach based on average linkage as cluster distance measure was chosen. The similarity calculation among all compound pairs was based on common occurring small molecule fragments. The fragment library was established by determining the largest common substructure between pairs of molecule fragments keeping ring systems intact. Each fragment had a minimum of four atoms. Pairwise compound similarities were calculated based on molecule fragment overlap using the tanimoto coefficient [192] between molecular fragment sets. All atoms were considered equivalent by element type. Bonds were compared by order (single, double, triple), ring membership, and aromaticity.

Calcium mobilization assay assesses ligands for mGlu5 modulation response

(This section was kindly provided by Alice L. Rodriguez)

HEK 293 cells expressing rat mGlu₅ were plated in black-walled, clear-bottomed, poly-D-lysine coated 384-well plates in 20 μ L of assay medium (DMEM containing 10% dialyzed FBS, 20 mM HEPES, and 1 mM sodium pyruvate) at a density of 20K cells/well. The cells were grown overnight at 37 °C in the presence of 5% CO₂. The next day, medium was removed and the cells incubated with 20 μ L of 2 μ M

Fluo-4, AM prepared as a 2.3 mM stock in DMSO and mixed in a 1:1 ratio with 10% (w/v) pluronic acid F-127 and diluted in assay buffer (Hank's balanced salt solution, 20 mM HEPES, and 2.5 mM probenecid) for 45 minutes at 37 °C. Dye was removed, 20 µL of assay buffer was added, and the plate was incubated for 10 minutes at room temperature. Compounds were serially diluted 1:3 in DMSO into 10 point concentration response curves and transferred to daughter plates using the Echo acoustic plate reformatter (Labcyte, Sunnyvale, CA) followed by further dilution into assay buffer to a 2x stock using a Thermo Fisher Combi (Thermo Fisher, Waltham, MA). Ca²⁺ flux was measured using the Functional Drug Screening System (FDSS7000, Hamamatsu, Japan). After establishment of a fluorescence baseline for about 3 seconds, the test compounds were added to the cells, and the response in cells was measured. 2.3 minutes later an EC₂₀ concentration of the mGlu₅ receptor agonist glutamate was added to the cells, and the response of the cells was measured for 1.9 minutes; an EC₈₀ concentration of agonist was added and readings taken for an additional 1.7 minutes. Data were collected at 1 Hz. Concentration response curves were generated using a four point logistical equation with XLfit curve fitting software for Excel (IDBS, Guildford, U.K.) or GraphPad Prism (GraphPad Software, Inc., La Jolla, CA).

Competition binding assay determines interaction of ligands with MPEP site

(This section was kindly provided by Alice L. Rodriguez)

The allosteric antagonist MPEP analog (³H)methoxyPEPy was used to evaluate the ability of test compounds to interact with the MPEP site on mGlu₅. Membranes were prepared from rat mGlu₅ HEK 293 cells as previously described (Rodriguez et al., 2005). Compounds were serially diluted in DMSO then added to assay buffer (50 mM Tris/0.9% NaCl, pH 7.4) to reach a 5x stock and 50 µL test compound was added to each well of a 96 deep-well assay plate. 150 µL aliquots of membranes diluted in assay buffer (20 µg/well) were added to each well. 50 µL (³H)methoxyPEPy (2 nM final concentration) was added and the reaction was incubated at room temperature for 1 hour with shaking. After the incubation period, the membrane-bound ligand was separated from free ligand by filtration through glass-fiber 96 well filter plates (Unifilter-96, GF/B, PerkinElmer Life and Analytical Sciences, Boston, MA). The

contents of each well were transferred simultaneously to the filter plate and washed 3-4 times with assay buffer using a cell harvester (Brandel Cell Harvester, Brandel Inc., Gaithersburg, MD). 40 μ L scintillation fluid was added to each well and the membrane-bound radioactivity determined by scintillation counting (TopCount, PerkinElmer Life and Analytical Sciences). Non-specific binding was estimated using 5 μ M MPEP. Concentration response curves were generated using a four parameter logistical equation in GraphPad Prism (GraphPad Software, Inc., La Jolla, CA).

Discussion

Allosteric binding selectivity for receptors of the mGlu GPCR family is driven by ligand binding behavior at the receptors TM domain. With the discovery of a potential alternative allosteric mGlu₅ binding site in addition to the already known MPEP binding site new possibilities to investigate more selective allosteric modulators for the mGlu₅ receptor are revealed. The goal of this study is to elucidate the chemical space of ligands interacting with this novel binding site. Thus, a series of CPPHA-like compounds was developed to investigate determinants of the alternative binding site further. The set of compounds contained 275 molecules comprised of 130 active and 145 inactive compounds that show structural similarity to CPPHA. In an effort to investigate the structural distinctiveness of these compounds a similarity analysis identified 17 clusters. Each cluster center is shown with its main scaffold and a percentage of involved active and inactive compounds (see Figure 5). A clear separation between clusters of solely active or inactive compounds was only given in clusters 3, 11, 14 and 4, 7, 10, 15, 16, respectively. This clustering scheme suggests a partially intertwined structural landscape for active and inactive molecules. Therefore, a pure structural separation is not possible.

On this basis, a QSAR model was trained on a data set consisting of the aforementioned 130 active compounds classified as active and 145 inactives complemented with all compounds available from a previous study [43] to identify mGlu₅ potentiators for the MPEP binding site (see Figure 6). QSAR models for the identification of mGlu₅ PAMs [43] and NAMs [183] were re-trained with the current version of BCL::ChemInfo. While re-training the QSAR models the quality measure results improved

substantially observing PPV rates of 51% (PAM) and 19% (NAM) indicating a substantial improvement compared to the published results when applying a similar FPR cutoff. These results are due to the implementation of regularization features to the ANN algorithm such as neuronal dropout [193], a technique that allows to train QSAR model with increased predictive generalization.

A third QSAR models was trained to identify allosteric binders for the mGlu₅ CPPHA site with an overall PPV of 0.73 when iteratively average until a FPR cutoff of 0.1%. When just evaluating the predictions for CPPHA-like actives and inactives in the independent data set a prediction accuracy of 79% was achieved. The strong predictive performance can be explained due to the extreme imbalanced character of the training data set and the distinctive structural features of the 130 active compounds in comparison to all categorized inactives. It is also expected that the QSAR model will suffer from the lack of cover chemical space of the active compounds. Thus, potential modulators interacting with the CPPHA site might be missed. The original QSAR model for mGlu₅ was trained to identify all modulators, mainly the one that bind the MPEP binding site. Since the set of actives was structurally different from the CPPHA analog series we assume these compounds to supplement the existing set of inactive compounds.

The three QSAR model were applied to virtually screen the small organic compound library *eMolecules* containing over 4 million molecules. A set of 4,719 molecules was predicted to be active for specific interaction with the CPPHA site, predicted inactive by the general mGlu₅ PAM QSAR model, and predicted inactive by the mGlu₅ NAM model. Medicinal chemistry filters, REOS and PAINS, reduced this set further down to only 103 compounds. From this pool of compounds 63 molecules were chosen by predicted activity, structural diversity, and acquisition cost.

A total of seven compounds were identified to express a biological response in mGlu₅. Six of these compounds were categorized as PAM and one as antagonist. The three most potent compounds reached a glutamate response of 40-50% of the maximal signal with potencies above 10 μ M. This relatively weak signal strength is an expected result given the primary goal of this study was to find new chemotypes

through scaffold-hopping that exhibit at least a weak response but interact selectively with the mGlu₅ site related to CPPHA binding.

To assess whether these seven compounds are indeed CPPHA site-specific binders further investigation of the location of binding was required. An experiment to determine the displacement of (³H)methoxyPEPy gave evidence that five of the identified seven compounds indeed interact with a different binding site distinct from the MPEP site.

That is an exciting result given the limited data set used to build the site-specific QSAR model. The five identified structures show a significant divergence from the reference CPPHA structure indicative of scaffold hopping through the involved QSAR modeling. With this approach demonstrates that the identification of target-specific binders is possible and is a valuable method in identifying novel lead compounds implicated in less adverse effects.

Conclusions

In summary, mGlu receptors of subtype 5 are suggested as possible candidates for drug targets involved in treatment of a range of CNS disorders. An alternative binding site distinct from the already known MPEP binding site was identified through mutagenesis and binding experiments involving residues F585 (TM1) and A809 (TM7) in the 7TM helices region. A comprehensive set of QSAR models was employed to query the commercially available compound library *eMolecules* for specific binders to mGlu₅ binding site related to CPPHA binding. A resulting set of 4,719 candidate molecules was identified. After the application of medicinal chemistry filters a subset of 63 molecules was chosen for compound acquisition. Subsequent calcium binding experiments revealed an increased biological response to mGlu₅ *in vitro* for 7 out of the ordered 63 compounds. Further experimental validation through displacement of the radio ligand (³H)methoxyPEPy confirmed 5 out of 7 compounds to not bind the MPEP site, indicating binding to an alternative mGlu₅ site associated with CPPHA binding. This corresponds to an approximated 8% (5/63 compounds) hit rate compared to the initial 0.09%.

These are very exciting results showing that QSAR modeling can identify novel chemotypes through scaffold hopping given only limited data sets. The confirmed compounds do not bind to the already known MPEP site, implying the aforementioned mGlu₅ site associated with CPPHA binding. The binding affinity can be refined through SAR optimization. As shown in this research virtual screening enables the prioritization of site-specific novel modulators, and thus contributes to the reduction of adverse effects of drug candidates. The here identified molecules have great potential to serve as novel lead or probe compounds in subsequent drug discovery campaigns.

5. SMALL MOLECULE PROPERTY PREDICTIONS

This chapter is based on publications [1] and [2].

Comparative analysis of machine learning techniques for the prediction of logP

Introduction

The process of modern drug design involves eliminating compounds with undesirable properties from the available chemical space while optimizing efficacy. The ability to predict properties which influence absorption, distribution, metabolism, and excretion of compounds prior to synthesis, such as the octanol-water partition coefficient (logP), could drastically reduce both the cost and time involved in drug discovery. Computational models can quickly assess the properties of large sets of compounds *in silico*. LogP, a measure of hydrophobicity or hydrophilicity of a molecule indicates whether a compound reaches a target protein as it influences the ability to cross the blood/brain barrier [194, 195]. It plays further a key role in the binding of a ligand to a target in aqueous solution [196]. Formally, logP is the logarithm of the equilibrium ratio of concentrations of a compound in the organic and aqueous phases of an octanol-water system [197]. LogP is a widely used, well-defined property with experimental values available for large numbers of compounds, which makes it ideal for prediction by machine learning methods.

A well-established method for prediction of logP is XlogP [198], which assigns each atom in the molecule an empirically-determined contribution depending on its type and then sums these contributions for the logP estimation of the entire molecule. This incremental method resembles a multiple linear regression model. We test the hypothesis that logP has a nonlinear dependence on composition, charge distribution, and shape of the molecule. Therefore, we expect non-linear models to improve prediction accuracy.

Machine learning techniques have been successful in approximating nonlinear separable data in Quantitative Structure Property Relationship studies [83, 130, 199, 200]. Here, we present several

predictive models for logP using machine learning techniques including artificial neural networks (ANN) [22], support vector machines with the extension for regression estimation (SVR) [201], and kappa nearest neighbors (kNN) [65].

Machine learning techniques

Artificial Neural Networks

The utility of ANNs for classification is well-known in chemistry and biology [202-205]. ANNs model the human brain and, thus, consist of layers of neurons linked by weighted connections w_{ji} . The input data x_i are summed according to their weights, activation function applied, and output used as the input to the j -th neuron of the next layer. For a three-layer feed forward ANN, such a training iteration would proceed as:

$$y_j = f(\text{net}_j) = f(\sum_{i=1}^d x_i w_{ji}) \quad 1 \leq j \leq n_H \quad (1)$$

$$z_k = f(\text{net}_k) = f(\sum_{j=1}^{n_H} y_j w_{kj}) \quad 1 \leq k \leq c \quad (2)$$

where $f(x)$ is the activation function, d is the number of features, n_H is the number of hidden neurons, and c is the number of outputs. For supervised training, the difference between the calculated output z_k and the target value t_k determines the errors for back-propagation:

$$\Delta w_{kj} = \eta(t_k - z_k) f'(\text{net}_k) y_j \quad (3)$$

$$\Delta w_{ji} = \eta[\sum_{k=1}^c w_{kj} (t_k - z_k) f'(\text{net}_k)] f'(\text{net}_j) x_i \quad (4)$$

The ANN training iterations produce weight changes that minimize the rmsd between the predicted and target values,

$$\text{rmsd} = \sqrt{\frac{\sum_{i=1}^n (\text{exp}_i - \text{pred}_i)^2}{n}} \quad (5)$$

which in this case is predicted and experimental logP values, respectively.

In this study, the ANNs have up to 1142 inputs, 8 hidden neurons, and one output (logP). The activation function of the neurons is the sigmoid function:

$$g(x) = \frac{1}{1+e^{-x}} \quad (6)$$

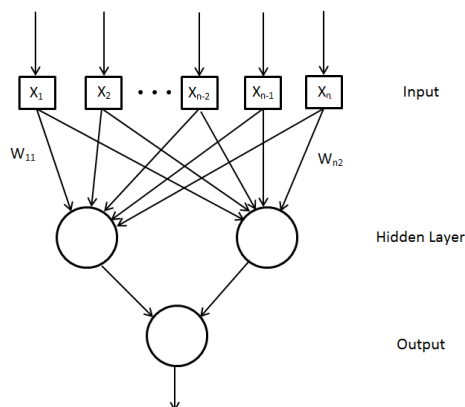


Figure 11: Schematic view of an ANN

Up to 1,142 descriptors are fed into the input layer. The weighted sum of the input data is modified by the activation function and serves as input to the next layer. The output describes the predicted logP value of the molecule. This figure was adapted from [1].

Support Vector Machines

The second machine learning approach applied in this study is SVM learning with extension for regression estimation [206, 207]. Linear functions defined in high-dimensional feature space [21], risk minimization according to Vapnik's ϵ - intensive loss function, and structural risk minimization [96] which minimizes the risk function consisting of the empirical error and the regularized term are the core principles integrated in SVM regression estimation.

The training data is defined by $(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$ with $i = 1, \dots, l$ where l is the total number of available input data pairs consisting of molecular descriptor data and experimental logP value. The following function defines a linear plane in a high-dimensional space for SVM estimation:

$$f(x, w) = w * \phi(x) + b \quad (7)$$

where $\phi(x)$ describes a nonlinear transformation function as a distance measure in an input space X . The parameter w describes a normal vector perpendicular to the separating hyperplane whereas b is a bias parameter. Both parameters are optimized by estimating the minimum of Vapnik's linear loss function as a measure of the error of approximation:

$$|y - f(x, w)| = \begin{cases} 0, & \text{if } |y - f(x, w)| \leq \varepsilon \\ |y - f(x, w)| - \varepsilon, & \text{otherwise} \end{cases} \quad (8)$$

The error is zero if the difference between the measured value y and the predicted value $f(x, w)$ is less than a given threshold ε . Thus, Vapnik's insensitivity loss function defines an ε - tube (Figure 12).

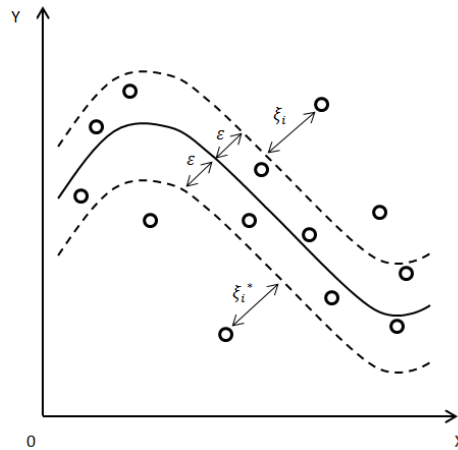


Figure 12: Schematic view of a Support Vector ε - tube.

Data points in ε - tube are not penalized, while points outside the tube get a penalty according to their distance from the tube edge. This figure was adapted from [1].

Predicted values positioned within the ε - tube have an error of zero. In contrast, data points outside the tube are penalized by the magnitude of the difference between the predicted value and the outer rim of the tube. The regression problem is solved by minimizing function L :

$$L_{w,\xi,\xi^*} = \frac{1}{2} \|w\|^2 + C(\sum_{i=1}^l \xi_i + \sum_{i=1}^l \xi_i^*) \quad (9)$$

under constraints:

$$\begin{aligned} y_i - g(x, w) &\leq \varepsilon + \xi_i \\ g(x, w) - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i^{(*)} &\geq 0 \quad i = 1, \dots, l \end{aligned} \quad (10)$$

where the slack variables ξ_i and ξ_i^* are shown in Fig 2 for measurements above and below an ε - tube, respectively. Both slack variables are positive values and their magnitude can be controlled by penalty parameter C . To estimate a numerical solution the optimization problem is converted into the given dual problem by:

$$\begin{aligned} f(x) &= \sum_{i=1}^{N_{SV}} (\alpha_i - \alpha_i^*) * K(x_i, x) + b \\ 0 &\leq \alpha_i \leq C, \\ 0 &\leq \alpha_i^* \leq C \end{aligned} \quad (11)$$

where α_i and α_i^* define Lagrange multipliers associated with ξ_i and ξ_i^* , N_{SV} shows the number of support vectors SV defining the SVM and $K(x_i, x_j)$ denotes the kernel function. In this study the Radial Basis Function kernel

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\gamma^2}} \quad (12)$$

was used to train the Support Vector Machine. The influence of the approximation error and the weight vector $\|w\|$ norm is balanced by the penalty constant C . It is optimized along with kernel parameter γ by using a grid search approach on a monitoring dataset.

Kappa Nearest Neighbors

The third machine learning approach utilized in this research is the kNN [65, 208-211]. kNNs are considered an unsupervised learning algorithm. This method uses a distance function to calculate pairwise distances between query points and reference points, where query points are those to be classified (Figure 13). The predicted value of a query point is then that of the weighted average of its *kappa* nearest reference points. In this research, the distance measure was the Euclidean distance between feature vectors:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (13)$$

The reference activities were weighted as $\frac{1}{\text{distance}}$, and the value of *kappa* was 5.

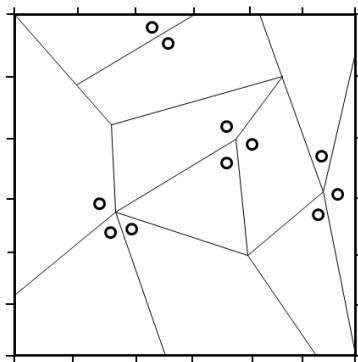


Figure 13: Schematic view of kNN cluster centers with its determined nearest neighbor environments. This figure was adapted from [1].

Training data

The octanol-water partition coefficient, or partition constant, is a measure of differential solubility of a substance. Specifically, it is the ratio of concentrations of a substance in the two phases of a mixture of the two immiscible solvents. These partition constants are useful in approximating the distribution of a substance within the body in medicinal settings. In pharmacology, logP is an indicator of how easily a substance can reach its intended target, and influences absorption, distribution, metabolism, and excretion

(ADME) properties of the substance. Thus, certain ranges of logP are desirable depending on the intended target and interaction of the substance making logP an important property in drug discovery. The ability to accurately predict this property is beneficial in the early stages of therapeutic design, such as during virtual high-throughput screening efforts and in analogue prioritization and optimization.

The training data for this investigation was obtained through data mining of the MDL Drug Data Report (MDDR) [212] and Reaxys [213] databases, as well as through literature searches using SciFinder [214]. Data mining resulted in ~26,000 compounds with experimentally determined values of logP. Of the compounds retrieved, the values of logP ranged from approximately -14 to 14. From this range, 13% of the compounds were removed from the training set from both extrema reaching a range of -5 to 8 in order to eliminate possible outliers at the limits of logP determination.

The remaining molecules in the training data set were numerically encoded using a series of transformation-invariant descriptors which serve as unique fingerprints.

Implementation / Method

All machine learning algorithms, and descriptor calculations used for this study were implemented in our in-house C++ class library, the BioChemistryLibrary (BCL). A third-party 3D conformation generator, CORINA [12], was used to generate 3D coordinates for the molecules prior to descriptor calculation.

Dataset generation

The data set used in this study was obtained through data mining and filtering which resulted in a final data set of 22,582 molecules. During the training of the models, 10% of the data set was used for monitoring and 10% were used for independent testing of the trained models, leaving 80% for the training data set.

Quality measure

The machine learning methods are evaluated by calculating the rmsd (eq. 5) using the cross-validated models.

The average rmsd of the cross-validated models for a feature set with n features is used to determine the predictive ability of the model. Additionally, correlation plots and the calculation of r^2 are also used to evaluate the trained models.

$$r^2 = \left(\frac{n \sum (exp \cdot pred) - \sum exp \sum pred}{\sqrt{[n \sum (exp^2) - (\sum exp)^2][n \sum (pred^2) - (\sum pred)^2]}} \right)^2 \quad (14)$$

Feature selection

1142 descriptors in 54 categories were generated using the BCL. The 54 categories consisted of scalar descriptors, as well as 2D and 3D autocorrelation functions, radial distribution functions, and van der Waals surface area weighted variations of each of the non-scalar descriptors (see Table 6).

Table 6: The original molecular descriptors by category

	Descriptor Name	Description	
Scalar descriptors	Weight	Molecular weight of compound	
	HDon	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen	
	HAcc	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in molecule	
	TPSA	Topological polar surface area in Å ² of the molecule derived from polar 2D fragments	
Vector descriptors	Ident	weighted by atom identities	
	2D Autocorrelation (11 descriptors) /	SigChg	weighted by σ atom charges
		PiChg	weighted by π atom charges
	3D Autocorrelation (12 descriptors) /	TotChg	weighted by sum of σ and π charges
		SigEN	weighted by σ atom electronegativities
	Radial Distribution Function (48 descriptors)	PiEN	weighted by π atom electronegativities
		LpEN	weighted by lone pair electronegativities
		Polariz	weighted by effective atom polarizabilities
	Every Vector descriptor available with and without van der Waals surface area weighting		

Sequential forward feature selection [215] was used for feature optimization for each machine learning technique individually. Additionally, each feature set was trained with 5-fold cross-validation. To cope

with the computational expense associated with this thorough feature selection process, a subset of the training data was used taking randomly only 8,000 of the 22,582 molecules. The number of models generated during this process for each training method was $\sum_{i=1}^{54} 5(n-1)$. Upon identification of the optimized feature set for each algorithm, any algorithm-specific parameters were optimized using the entire training data set and using 5-fold cross-validation.

Results

ANNs were trained using 7,500 iterations of simple propagation evaluating the rmsd every 100 steps during the feature optimization process. For the training of the final model with the optimized feature set, 100,000 iterations were performed evaluating the rmsd every 500 steps. The weight matrices were randomly initialized with values in the range of -0.1 to 0.1 if the rmsd of the monitoring data set had not improved in the last 10,000 iterations. The ANN algorithm runs on graphics processing units (GPUs) using OpenCL implemented within the BCL. The training time was 28 minutes per final network on a C2050 NVidia GPU on a Dell T3500 with 8-core Xeon 3.2 GHz microprocessor running CentOS 5 64-bit. For the best model, an rmsd of 1.20 for the independent data set was achieved.

SVMs were trained using a C of 1.0 and γ of 0.1 during the feature optimization process. Upon identification of the optimal feature set, the cost and γ parameters were optimized to 0.1 and 0.1, respectively. The training time for the final model was 12 minutes using 6 cores. Using these optimized parameters, the cross-validated model achieved an rmsd for the independent data set of 1.21.

The kNN algorithm was used to predict the logP values of the training, monitoring, and independent data sets. The value of kappa, the number of neighbors to consider, was optimized with the full data set using the optimized feature set determined during the feature selection process. The prediction time for the final model was 0.75 minutes on 8 cores. Using an optimal kappa of 5, the relative rmsd achieved for the independent data set was 1.03.

Table 7: Model statistics for best predictors

Method	RMSD	R ²
ANN	1.20	0.70
SVM	1.21	0.67
kNN	1.03	0.72
XlogP	1.41	0.56

In order to further evaluate the resulting models, cross correlation plots were created on the independent data sets of each model and compared with that of the XlogP algorithm [198].

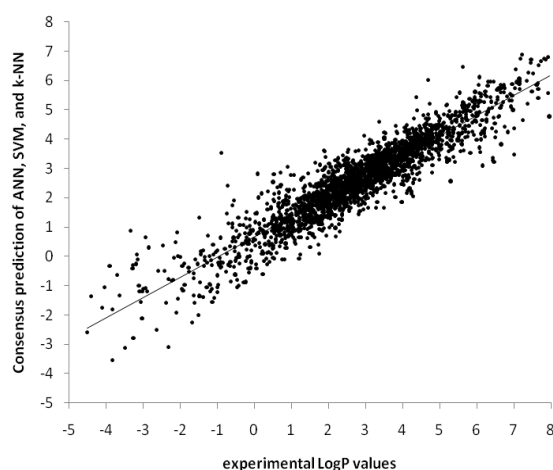


Figure 14: Correlation plot of best consensus predictions compared to experimental data.

ANN/SVM/kNN predictions are compared with Reaxys/MDDR experimental values. This figure was adapted from [1].

Consensus predictors were also created by averaging the predictions using different combinations of models. This yields better results with the best rmsd of 0.86 achieved using ANN/SVM/kNN models (Table 8).

Table 8: Consensus Predictors

Method	RMSD	R ²
ANN/SVM	1.04	0.80
ANN/kNN	1.06	0.81
SVM/kNN	0.99	0.77
ANN/SVM/kNN	0.86	0.86

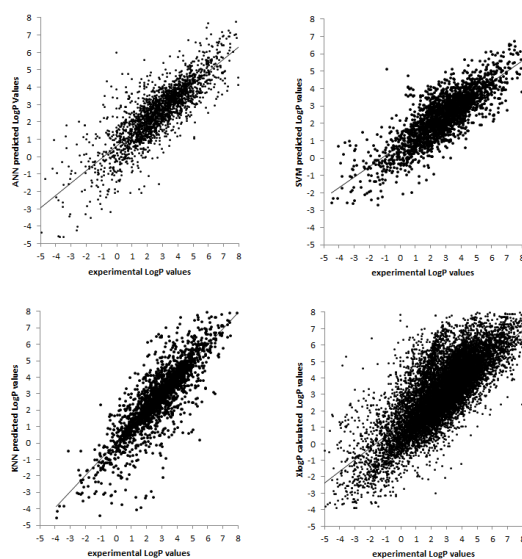


Figure 15: ANN, SVM, kNN, and XlogP predictions compared to Reaxys/MDDR experimental values. This figure was adapted from [1].

Conclusions

Here, we present the utility of a series of machine learning techniques for the construction of predictive models capable of predicting logP, the water-octanol partition coefficient, with high accuracy compared to XlogP. The XlogP algorithm is a standard and is used by the NIH-funded PubChem database as well as the Chemical Abstracts Service. We have shown that the three models using artificial neural networks, support vector machines, and kappa nearest neighbors outperform this method as do all of the consensus models.

The best individual model was found to be that of the kNN. This is likely due to the kNNs ability to predict only around the cluster space of kappa neighbors while the other machine learning methods take the entire sample space into account when training. This all-inclusive training allows the SVM and ANN to over-train on the mean due to underrepresentation of the extrema. This is reflected by the slope of the best-fit linear regression lines in Figure 14. The best performing method was the consensus prediction of ANN/SVM/kNN models which achieved an rmsd of 0.86. This is likely due to the ANNs and SVMs ability to predict the core range of the logP space more accurately while the kNN was more accurate at the extrema.

Quantitative structure property modeling for the prediction of DMPK parameters intrinsic clearance and plasma protein binding

Introduction

During the 1990's, poor pharmacokinetic and bioavailability properties accounted for approximately 40% of drug candidate attrition during human trials [216]. A decade later, these properties accounted for approximately 10% of attrition during human trials due to the implementation of early determination (pre-clinical) of drug metabolism and pharmacokinetics (DMPK) properties in the drug discovery workflow through both *in vitro* and *in vivo* studies. While many proposed new chemical entities (NCE) are now eliminated in earlier stages, these preliminary studies are time consuming and add to the mounting real and opportunity costs which now approach an estimated \$1.30 – \$1.76 billion [217, 218]. Thus, the use of *in silico* models for the prediction of these DMPK properties trained on existing data would increase the efficiency of the drug discovery process while mitigating the costs [219]. Indeed, computational models can quickly assess large data sets of proposed molecules for DMPK parameters [1].

Two important DMPK properties which are routinely determined in drug discovery are microsomal intrinsic clearance (CL_{int}) and plasma protein binding as the fraction of unbound compound (f_u). CL_{int} is a measure of metabolism primarily by cytochrome P-450 (CYP) enzymes in the vesicles of the smooth endoplasmic reticulum of hepatocytes. CYP enzymes contribute to the metabolism of approximately 75%

of the top 200 most prescribed drugs in the United States [220]. f_u is an indication of the extent to which a compound binds to plasma proteins which influences to a large degree pharmacokinetics, efficacy, and toxicology *in vivo* [221, 222].

Previous work has proven machine learning techniques useful in the approximation of nonlinear separable data in Quantitative Structure Property Relationship (QSPR) studies [1, 43, 83, 130, 200]. Here, we present several predictive models based on machine learning techniques for human and rat CL_{int} as well as human and rat f_u . The machine learning techniques used include artificial neural networks[50], support vector machine with the extension for regression[201], kappa nearest neighbor[65], and Kohonen networks[223].

Methods

All descriptors calculated and machine learning algorithms used in this study are implemented in the in-house C++ class library, the BioChemistry Library (BCL). CORINA, a 3rd-party 3D conformation generator, was used for the generation of 3D coordinates prior to descriptor calculations [12].

Machine learning techniques

Artificial Neural Network

The utility of artificial neural networks (ANN) for classification is proven in chemistry and biology [202-205]. ANNs model the human brain, consisting of layers of nodes linked by weighted connections w_{ji} . Input data x_i are summed by their weights, followed by the application of an activation function, and the output used as the input to the j -th neuron of the next layer.

For a common feed forward ANN with a single hidden layer, the training iteration would proceed as:

$$y_j = f(\text{net}_j) = f(\sum_{i=1}^d x_i w_{ji}) \quad 1 \leq j \leq n_H \quad (1)$$

$$z_k = f(\text{net}_k) = f(\sum_{j=1}^{n_H} y_j w_{kj}) \quad 1 \leq k \leq c \quad (2)$$

where $f(x)$ is the activation function (commonly sigmoid), d is the number of features, n_H is the number of hidden neurons, and c is the number of outputs. The difference between the calculated output z_k and the target value t_k , provides the errors for back-propagation through the network:

$$\Delta w_{kj} = \eta(t_k - z_k) f'(net_k) y_j \quad (3)$$

$$\Delta w_{ji} = \eta \left[\sum_{k=1}^c w_{kj} (t_k - z_k) f'(net_k) \right] f'(net_j) x_i \quad (4)$$

The weight changes produced attempt to minimize the objective function (RMSD) between the predicted and target values,

$$RMSD = \sqrt{\frac{\sum_{i=1}^n (exp_i - pred_i)^2}{n}} \quad (5)$$

In this study, the ANNs have up to 1284 inputs, 8 hidden neurons, and one output (DMPK parameter of interest). The activation function of the neurons is the sigmoid function:

$$g(x) = \frac{1}{1+e^{-x}} \quad (6)$$

Support Vector Machine with extension for regression estimation

SVM learning with extension for regression estimation [201] represents a supervised machine learning approach successfully applied in the past [1, 43, 200]. The core principles in SVR lay in linear functions defined in high-dimensional feature space [21], risk minimization according to Vapnik's ϵ - intensive loss function, and structural risk minimization [224] of a risk function consisting of the empirical error and the regularized term.

The training data is described by $(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$ with $i = 1, \dots, l$ where l is the total number of available input data pairs consisting of molecular descriptor data and the experimental DMPK property.

Given a defined error threshold ε , SVR seeks to approximate Vapnik's insensitivity loss function through the definition of an ε - tube incorporating all data points of the given problem. The error is zero if the difference between the experimentally measured value and the predicted value is less than ε .

Predicted values positioned within the ε - tube have an assigned error value of zero. On the opposite, data points outside the tube are penalized by the distance of the predicted value from the edge of the tube. The solution to the regression problem is obtained by minimizing the following function L :

$$L_{w,\xi,\xi^*} = \frac{1}{2} \|w\|^2 + C(\sum_{i=1}^l \xi_i + \sum_{i=1}^l \xi_i^*) \quad (7)$$

under constraints:

$$y_i - g(x, w) \leq \varepsilon + \xi_i \quad , \quad g(x, w) - y_i \leq \varepsilon + \xi_i^*$$

$$\text{and } \xi_i^{(*)} \geq 0 \quad i = 1, \dots, l$$

where the parameter w describes a normal vector perpendicular to the separating hyperplane in the higher dimensional space. The slack variables ξ_i and ξ_i^* are for measurements above and below an ε - tube, respectively. Both slack variables are positive values and their magnitude can be controlled by the penalty parameter C . In this study the Radial Basis Function kernel

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\gamma^2}} \quad (8)$$

was applied as distance measure. The penalty constant C determined the influence of the approximation error penalty. A grid search approach was conducted to optimize both parameters γ and C using a monitoring dataset.

Kappa Nearest Neighbor

Kappa nearest neighbor (kNN) was also utilized in this study [65, 208-211]. kNNs are an unsupervised learning algorithm using a distance function to calculate pair-wise distances between query points and

reference points. Query points are those to be classified. The query point is classified through a weighted average of the known output of its *kappa* nearest reference points. The distance measure used in this study was the Euclidean distance measure between feature vectors:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (9)$$

The reference activities were weighted as $\frac{1}{d(x,y)}$, and the value of *kappa* was optimized for each data set.

Kohonen Network

The kohonen network represents an unsupervised learning algorithm. It is conceptually derived from artificial neural networks consisting of one input layer connected by weighted connections with a two dimensional grid of neurons, the kohonen network [225].

The training data defined by pairs of numerical molecular descriptor data x_i and the respective experimental DMPK parameter y_i ($x_i \in X \subseteq R^n, y_i \in Y \subseteq R$) with $i = 1, \dots, l$ where l is the total number of available input data pairs.

For every training data point, a node most similar to the data point is determined for placement in the grid. Weight vectors are updated using the Gaussian kernel as a neighborhood function. A radius of four neighboring nodes is considered.

To determine the classification result of an unknown compound, the most similar node is determined and the average prediction values of all neighboring nodes are computed.

Data set generation

The data sets used in this study were obtained through the Vanderbilt Center for Neuroscience Drug Discovery. The data sets ranged from 386 (human *fu*) to 601 (rat CL_{int}) as seen in Table 9.

Table 9: Data set composition

Data Set	Number Molecules
Rat <i>fu</i>	388
Human <i>fu</i>	386
Rat CL _{int}	601
Human CL _{int}	576

Three-dimensional conformations for the molecules were generated. The molecules in the data sets were then numerically encoded using transformation-invariant descriptors (Table 10) which represent features for the machine learning techniques.

Quality measures

The calculated RMSD (eq. 5) is used to evaluate the predictive power of the machine learning models. Specifically, the average RMSD of the cross-validated models for a feature set with **n** features is used. The Pearson, R_p^2 , and Spearman, R_s^2 , correlation coefficients are computed.

$$r_{P/S} = \frac{n \sum(\text{exp} * \text{pred}) - \sum \text{exp} \sum \text{pred}}{\sqrt{[n \sum(\text{exp}^2) - (\sum \text{exp})^2][n \sum(\text{pred}^2) - (\sum \text{pred})^2]}} \quad (10)$$

Additionally, the Spearman correlation coefficient is defined as the Pearson (10) correlation coefficient between the rankings of variables.

Feature selection

The BCL was used to generate 1284 descriptors in 60 categories. The 60 categories consist of scalar, 2D and 3D autocorrelation functions, radial distribution functions, and van der Waals surface area weighted variations of each of the non-scalar descriptors (see Table 10).

Sequential forward feature selection [215] was used for feature optimization for each machine learning technique individually. Each feature set was trained with 5-fold cross-validation. The number of models generated during this process for each training method was $cv * \frac{n(n+1)}{2}$ where **n** is the number of feature categories and *cv* is the number of cross validations. Thus, 9150 models were trained for each machine learning algorithm on each data set during feature selection. Upon identification of the optimized feature

set for each algorithm, algorithm-specific parameters were then optimized using a grid search with 5-fold cross-validation.

Table 10: Molecular descriptors by category

	Descriptor Name	Description	
Scalar descriptors	Weight	Molecular weight of compound	
	HDon	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule	
	HAcc	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule	
	TPSA	Topological polar surface area in Å ² of the molecule derived from polar 2D fragments	
Vector descriptors	Ident	weighted by atom identities	
	2D Autocorrelation	SigChg	weighted by σ atom charges
	(11 descriptors) /	PiChg	weighted by π atom charges
	3D Autocorrelation	TotChg	weighted by sum of σ and π charges
		VCharge	weighted by VCharge atom charges
	(12 descriptors) /	SigEN	weighted by σ atom electronegativities
	Radial Distribution	PIEN	weighted by π atom electronegativities
	Function (48 descriptors)	LpEN	weighted by lone pair electronegativities
		Polariz	weighted by effective atom polarizabilities
	All molecular fingerprints are considered with and without van der Waals surface area weighting		

Results

During feature selection, ANNs were trained for 100 epochs of simple back-propagation using $\eta = 0.1$ and $\alpha = 0.5$ with weight updates and the evaluation of RMSD every step using 5-fold cross validation. Weight matrices were initialized randomly with values in the range [226]. A grid search was performed to optimize eta and alpha parameters using the optimized feature set and trained 100 epochs using 5-fold cross validation (Table 11). Eighty percent of each data set was used as the training set while 10% was used for the monitoring data set and 10% for the test data set.

SVMs were trained using a C of 0.1 and γ of 0.5 during the feature optimization process. Upon identification of the optimal feature set, the cost and γ parameters were optimized using a grid search approach (Table 11).

Table 11: Optimized parameters

Data Set	ANN (η, α)	SVM (C, γ)	KNN (k)
Rat <i>fu</i>	0.25/0.015625	2.0/0.25	5
Human <i>fu</i>	0.125/0.5	2.0/0.03125	24
Rat CL _{int}	0.03125/0.03125	0.25/0.25	14
Human CL _{int}	0.03125/0.0625	1.0/0.03125	16

The SVMs were trained 100 iterations and used 5-fold cross validation. Each iteration step accumulated up to 200 support vectors.

Table 12: Model correlation results ($r_{p/s}/\text{rmsd}$) for independent validation set

Machine Learning Method(s)	Rat	Human	Rat	Human
	<i>fu</i>	<i>fu</i>	CL _i	CL _{int}
ANN	9.52	8.47	1.08	1.10
ANN/KNN	9.53	8.60	1.09	1.18
ANN/KNN/Kohonen	9.29	8.78	1.08	1.17
ANN/KNN/Kohonen/SVM	8.79	8.83	1.01	1.10
ANN/KNN/SVM	8.73	8.70	0.99	1.08
ANN/Kohonen	9.17	8.78	1.07	1.13
ANN/Kohonen/SVM	8.39	8.78	0.97	1.04
ANN/SVM	8.00	8.51	0.92	0.97
KNN	9.43	8.67	1.09	1.23
KNN/Kohonen	9.07	8.89	1.07	1.19
KNN/Kohonen/SVM	8.36	8.84	0.98	1.10
KNN/SVM	7.98	8.58	0.94	1.06
Kohonen	8.56	8.94	1.05	1.15
Kohonen/SVM	7.38	8.65	0.91	1.01
SVM	4.35	7.49	0.71	0.82

KNNs were trained by optimizing kappa, the number of neighbors to consider, during feature selection from $k=1$ to $k=25$ using 5-fold cross validation (see Table 11). Eighty percent of each data set was used as reference features while 10% were queried as the monitoring data set. The remaining 10% was then used as a test set.

Kohonen networks were trained using a grid of 10x10 nodes. An ensemble model approach was then investigated by using all combinations of optimized models to arrive at single best predictors for each

data set. This approach provided the best models in terms of correlation. The resulting $\frac{r_p}{RMSD}$ values are listed in Table 12 with correlations plots shown for each of the best predicting models.

Conclusion

Here, we present ensemble models based on machine learning techniques capable of predicting several parameters relevant to drug discovery. We have shown that ensemble models are in some cases capable of

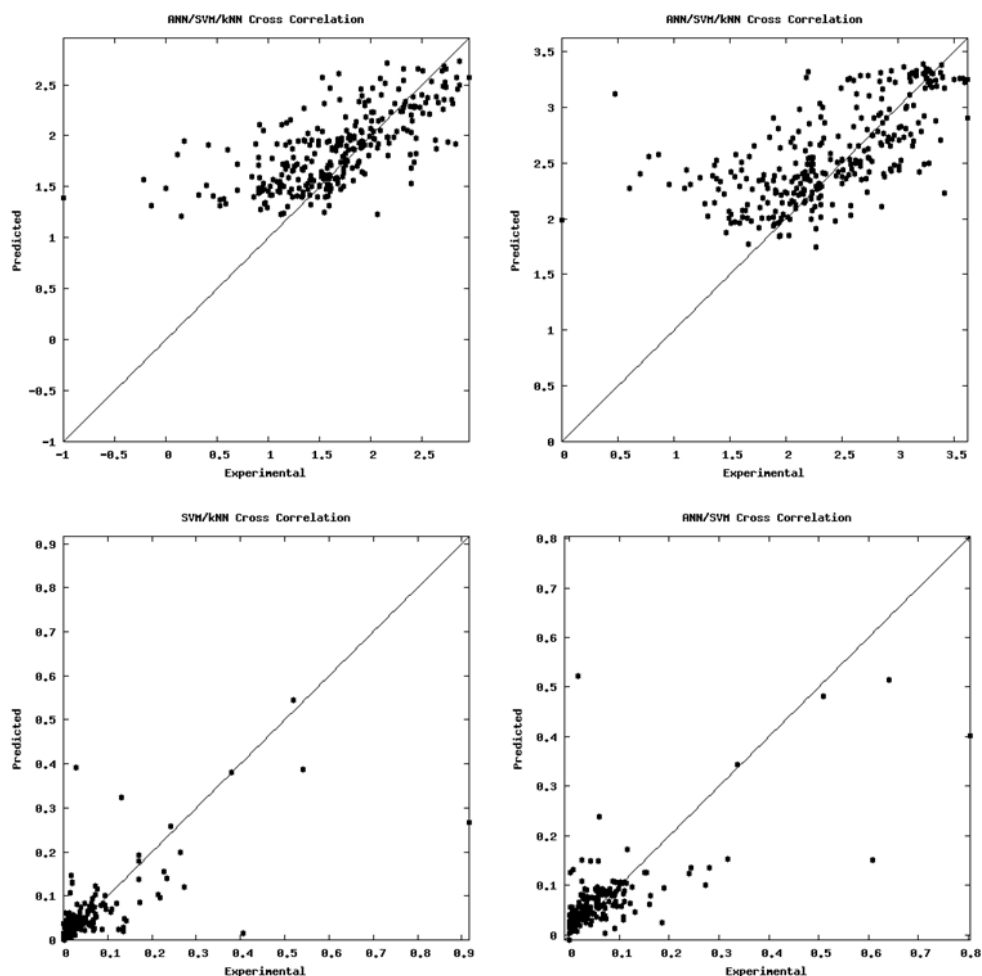


Figure 16: Overview of correlation plots for respective consensus models
 Correlation plots are shown for the machine learning models with the highest predictive power as determined by $\frac{r_p}{RMSD}$ for human CL_{int} , rat CL_{int} , human fu , and rat fu , respectively. The selected models are shown in bold in Table 12. This figure was adapted from [2].

outperforming single algorithms using artificial neural networks, support vector machines, kappa nearest neighbors, and kohonen networks.

KNN algorithm consistently performs very well for all 4 data sets examined. The predictors constructed during this study compare favorably against recent studies [227] and are of great utility in early drug discovery. The top scoring predictors for human and rat CL_{int} , and human and rat fu are KNN, ANN/KNN, Kohonen, and ANN/KNN, respectively.

6. DISCUSSION

Conclusions and future directions

The main focus of this thesis was to develop a virtual screening framework that allows for small molecule data handling, QSAR model development, optimization and analysis, as well as subsequent QSAR model prediction through virtual screening, an integral part of drug discovery projects. Two main goals were pursued. The first objective was the method development and benchmarking of BCL::ChemInfo as a novel and robust tool for virtual screening through QSAR modeling. The second objective was to assess whether the developed method is applicable to real-world scenarios in drug discovery. Different biological targets such as mGlu₅ modulation involved in schizophrenia, β -hematin crystallization inhibition involved in malaria, or molecular property prediction were highlighted in previous chapters. The present chapter will summarize key findings of each previous project and give suggestions on future perspectives.

Methods development and benchmarking of Quantitative Structure Activity Relationships

When designing a benchmark for QSAR modelling one of the critical challenges is to identify an array of data sets that are suitable for the task and yet diverse enough to span a wide range of applicability. Here, nine diverse protein targets spanning a variety of real-world drug targets were chosen from the PubChem database as benchmark data set. Biological assays are available for different stages of HTS. Primary screens evaluate an initial library of compounds for activity, typically in singlicate to reduce cost. Resulting hit compounds are typically investigated further by counter and confirmatory screens, typically in duplicate or triplicate. It was of essence to understand the validation hierarchy of each of the nine data sets to compile a high-quality benchmark set.

BCL::ChemInfo incorporates an array of different machine learning algorithms into one framework. Machine learning algorithms include ANNs, SVMs, KNs, kNNs, and DTs which can be trained in an

extensive cross-validated fashion. It is easily extendable to host future machine learning algorithms and unifies data set creation as well as handling for each machine learning algorithm.

The nature of biological assay data poses a further challenge. The ratio of active to inactive compounds determined in an HTS experiment is usually highly imbalanced. Inactive compounds are overrepresented by a factor of 100 to 1,000. Here, an appropriate alternative quality measure was proposed based on the ROC curve derivative TNR-TPR curve to evaluate machine learning models. This measure does not suffer from the imbalanced character of HTS assay data sets, and thus appropriately represents the performance of machine learning techniques.

Descriptor selection plays a significant role QSAR model optimization. The benchmark in this study confirmed that each combination of biological target and machine learning algorithm favor a different molecular descriptor set. A further improvement which was not part of this benchmark could be the introduction of a truly cross-validation wide independent data set or blind data set. Current descriptor selection schemes choose descriptor subsets based on the monitoring and training data partition while evaluating the machine learning technique on the independent data portion. The general reasoning is that the chosen descriptor set is representative enough for the chemical space of the data set. But different cross-validation iterations use the same independent data portion for descriptor selection and could introduce a chemical space bias. Therefore, a blind data set could circumvent this potential bias could in future studies.

Consensus prediction is another feature of BCL::ChemInfo. Through the consensus of different predictive outputs it is possible to improve the prediction accuracies of trained QSAR models. Our benchmark study showed a strong trend towards multiple models typically outperforming single predictors.

Future directions of this research involve the implementation of new machine learning algorithms. Deep learning approaches such as Restricted Boltzmann Machines (RBM) or Stacked denoising Autoencoders

(SdA) are of interest and showed promising results in research fields such as image recognition. The description of a proposed research project regarding SdAs is located in the Appendix.

A second future direction might be the incorporation of protein sequence data of the respective biological protein target. LB-CADD methods do not rely on the structural information of the target but have access to the respective amino acid sequence. This additional information has the potential to introduce novel discriminative features for QSAR model training. A proposed research project description regarding the combination of protein sequence and small molecule descriptors is located in the Appendix.

Discovery of pathway specific antimalarial hit compounds to diminish *P. falciparum*

The pharmaceutical companies GlaxoSmithKline (GSK) and Novartis tested combined approximately 3.7 million compounds against *P. falciparum*. Out of the respective libraries of compounds approximately 13,500 (GSK) and 5,700 (Novartis) hit compounds were identified to eliminate the malaria parasite. A collaborative study investigated an HTS effort at Vanderbilt University to elucidate compounds that kill the parasite by acting through a specific pathway inhibiting a vital crystallization process involving β – hematin, a synthetic form of hemozoin. From a screened set of 13,229 compounds, a set of 530 compounds emerged as inhibitors of β – hematin crystallization. This data was used as a knowledgebase to develop QSAR models with BCL::ChemInfo. This approach prioritized a subset of 249 GSK and 37 Novartis compounds with predicted activity against β – hematin formation. A subsequent experimental validation of the GSK predictions confirmed 40 out the predicted 249 compounds for β – hematin formation, resulting in a hit rate of 16% compared to an initial hit rate of 0.37%. A subset of 17 compounds was chosen to validate the fatal effect on the parasite and 15 out of 17 were confirmed. With the application of BCL::ChemInfo it was possible to screen *in silico* a larger chemical space than would be possible with conventional HTS. Current antimalarial drugs are becoming more and more ineffective while the drug resistance of the malaria parasite increases. A possible future direction for this project could be the incorporation of a variety of QSAR models trained against a range of *P. falciparum* targets.

This aspect would allow creating a profile prediction of future antimalarial molecules that are selective against specific pathways of action. PubChem and ChemBL provide sources of suitable HTS assay data.

Selective allosteric modulators for distinct mGlu₅ site related to CPPHA binding

Previous studies provided evidence for an alternative binding site other than the already known MPEP binding site for mGlu₅ modulation [168]. The positive allosteric modulator CPPHA is involved in binding exclusively to this alternative binding site. A series of CPPHA like compounds was prepared as the basis for *in silico* virtual screening with BCL::ChemInfo. In addition, data from previous research [84, 183] was pooled to develop three distinct QSAR model predicting unspecific positive allosteric modulation (PAM) for mGlu₅, unspecific negative allosteric modulation (NAM) for mGlu₅, and selective PAMs for mGlu₅ specific to the aforementioned alternative binding site. A virtual screening campaign was initiated to prioritize the compound library *eMolecules* containing approximately 4 million molecules. A total of 63 molecules were chosen after the application of medicinal chemistry filters. Preliminary experimental validation verified 6/63 compounds with weak general PAM activity for mGlu₅ resulting in an experimental hit rate of ~10%. Further testing for MPEP binding competition has to verify whether these 6 compounds exhibit selective binding. This project highlights the capabilities of BCL::ChemInfo for tailored drug discovery campaigns searching for specific modulators. Even though the discovered compounds exhibit weak binding for mGlu₅ the identified compounds can act as novel starting points for later SAR optimization for selective binding.

Future perspectives for this project could include the exploration of SB-CADD methods with the ROSETTA::LIGAND software package [228, 229]. A crystallographic model of mGlu₁ was recently published [230] which is closely related to mGlu₅. This SB-CADD approach could be applied as a *in silico* filter to confirm ligand binding prior to compound acquisition.

Molecular property predictions

The prediction of small molecule properties is an important application when evaluating molecule for drug metabolism and pharmacokinetics (DMPK) properties. This chapter is based on two research projects: development of a QSPR model for 1) the water-octanol partition coefficient (logP) prediction and 2) microsomal intrinsic clearance (CL_{int}) and plasma protein binding as the fraction of unbound compound (fu). Each property is essential for prospective drug candidates when interacting with the human body. In both studies BCL::ChemInfo was applied to develop a consensus QSPR model. The best consensus predictor was compared to the current gold standard for logP prediction (XlogP). A root mean square distance of 0.86 was achieved by BCL::ChemInfo compared to 1.21 for XlogP, a significant improvement. The second study investigated consensus prediction performance the properties CL_{int} and fu (human and rat) in a similar fashion as for logP. It was shown that ensemble models are capable of outperforming single algorithms using artificial neural networks, support vector machines, kappa nearest neighbors, and kohonen networks. Predictors constructed during this study compare positively against recent studies and are of great utility in early stage drug discovery campaigns.

Future perspectives for QSPR modeling with BCL::ChemInfo consider an expansion of respective property training data. The current data set used in this chapter is fairly small which makes extrapolation of extreme outlier compounds difficult for the constructed QSPR models.

Future perspectives of LB-CADD

With access to more HTS data challenges will emerge over time in the research field of LB-CADD. As pointed out in Chapter II, HTS data set curation is imperative to ensure high quality QSAR modeling. A future extension of BCL::Cheminfo could incorporate an automated evaluation of available HTS assay data for compound repositories such as PubChem, or ChEMBL. Related biological assay data sets could be identified for each repository and be cross-referenced for association with the same protein target. With minimal supervision by the user this approach would allow a compilation of primary and

confirmatory screens more easily and ensure a highly curated HTS data set well suited for QSAR modeling.

Not only will the number of available compounds increase the dimensions and thus the size of future QSAR training data sets but it will also encourage the develop of novel small molecular descriptors. Memory requirements will be a concern. There is the possibility that future small molecule training data sets will not fit into memory of a single computer anymore. What would be an efficient way to deal with this growing amount of QSAR training data? One possible strategy could encompass a distributed database approach with a network of multiple dedicated computers. Data set access could be coordinated by a central database server with access to all involved sub-databases. The compound information can be stored on multiple file systems with enough storage capacity and ensure a user friendly single point access to retrieve compound data. The challenging aspect would be adapting current machine learning algorithms to retrieve their training data in chunks rather than having access to the entire training dataset. In addition, it would be expected that the latency of data access will suffer due to the network communication overhead. An alternative coping strategy would be the clustering of molecules to ensure a maximum coverage of chemical space. The resulting reduction of compounds will ensure a manageable size of training data sets fitting into the memory of a single computer.

Machine learning is a progressing field of research. With adapting machine learning algorithms for LB-CADD one important question will arise: What is the best way to identify novel machine learning algorithms suitable for application in CADD? Recently emerging machine learning algorithms, such as deep belief networks based on Restricted Boltzmann machines and stacked denoising Autoencoders, or manifold learning algorithms like Isomap showed great success in applications such as image and voice recognition, sentiment analysis, or dimensionality reduction. Yet, benchmark studies have to show whether the prediction performance will hold up with competing algorithms when predicting the biological activities of small molecules for novel protein targets. To assess the algorithm performance Chapter II introduced an extensive benchmark of machine learning algorithms for biological datasets. It

can be assumed that with increasing quantity and quality of small molecule data the prediction performance of machine learning algorithms will improve simply by sampling a larger chemical space more precisely.

With the larger spectrum of available HTS data it is reasonable to expect an increase in the number of protein targets. QSAR modeling will not just focus on single targets, but models will be trained for an array of targets associated with entire signal transduction pathways. Predictions for a range of biological targets will foster a better understanding of the path of action for each screened compound. QSAR models can then be optimized to predict a compound profile assessing multiple targets simultaneously and giving insight to protein target interaction specificity of the ligand. The BCL::Cheminfo software suite is designed to accommodate this use case scenario.

Over time, it is expected that new classes of drug targets will be discovered that potentially require more tailored descriptors to model this new chemical problem. How can existing LB-CADD methods be adapted for these new drug targets? One example is protein-protein interactions. Ligand descriptors coupled with protein sequence information are one possible approach to incorporate this type of information into QSAR model training. Combining small molecule and protein sequence information descriptors would have the advantage of creating an association between determinants of both descriptor types. This encoding paradigm could be applicable for the encoding of biopharmaceuticals or simply ‘biologics’ in association with a small molecule of interest. Most biologics are complex molecules or mixtures of molecules extracted from living systems or produced by recombinant DNA technology. Examples are antibodies, vaccines, blood tissues, or somatic cells. Further, adding nucleic acid information could allow modeling of transcription factors associated with nucleic acid – protein interactions as drug targets. The encoding of DNA or RNA related information along with an associated small molecule- or protein-ligands is another approach to associate different types of targets and thus foster modeling of novel ligand interaction patterns. Given the development of appropriate descriptors

tailored for these drug targets BCL::ChemInfo can be easily adapted to handle the different modeling scenarios.

Interesting application possibilities could arise for LB-CADD modeling when longitudinal transcriptome data of assay experiments becomes available. The transcriptome of a cell changes dynamically dependent on its accessible genome defined through mechanisms like DNA methylation pattern or histone modification. With the rise of genetic sequencing technologies, a procedure called ‘whole transcriptome shotgun sequencing’ can reveal which RNA type is present and its quantity at a given point in time. Biological assay data could be collected at specific time intervals and biological effects of candidate compounds can be quantified. If assay experiments require cells to be lysed and thus destroyed to obtain a measurement the experiment could be repeated and its duration increased until the final evaluation. LB-CADD models could be trained on small molecule ligands associated with RNA representations and quantities. A possible application could be the detailed elucidation of candidate compound toxicity through predictions of the hit compound activity in relation to the environmental changes of the cell.

In general, research fields will become more synergistic. The fields of medicine, pharmacology, and genetics will progress to include more initial computational guidance. The acceleration of technological infrastructure will see a decrease in associated costs while more computational processing resources will be available. Virtual screening methods in LB-CADD will allow for quicker, more cost-effective development of highly specialized drugs that will interact with the research field of pharmacogenetics to see a realistic rise in personalized medicine. Currently, physicians prescribe medication with recommended dosages determined by clinical trials or previous experiences with the drug. However, the optimal medication dosage ranges widely from patient to patient and can be harmful or even deadly when too high. Thus, a future application of QSAR/QSPR modeling could entail the incorporation of small molecule information of the candidate drug, protein target information, and the associated genetic markup of a patient to predict likely appropriate dosage ranges suitable for the patient.

The failure rates of drug candidates passing clinical trials and being approved by the US Food and Drug Administration (FDA) is very high. A grand challenge for future drug discovery research will be to understand why certain drug candidates are unsuccessful in the clinic. Gathering this understanding will guide future drug discovery efforts to omit previous shortcomings. The prediction of pharmacokinetic properties of compounds will play a more substantial role in this challenge. Chapter V gave insight into the prediction of small molecule properties as part of ADMET (absorption, distribution, metabolism, excretion, and toxicology) modeling. An increased understanding of the effect of a drug candidate on the human body will ultimately determine the success of a drug candidate in clinical trials.

Another future challenge will be the acceptance of novel LB-CADD tools in the medicinal chemistry community. The vision is that technology will disperse. Virtual high-throughput screening will become a routine in drug discovery campaigns. How can this technology be efficiently brought to medicinal chemistry? To see such a trend in the future one can envision the following scenarios. Novel LB-CADD algorithms will have to be developed not only focusing on performance but also usability. User interfaces will need to be improved to be more self-explaining and user friendly. Another challenge is to educate researchers in the scientific community about the available LB-CADD methods and the benefit the computational tool can have for their research. To disseminate the method functionality, workshops can be developed to allow for an in-person experience. Towards this goal, a virtual high-throughput workshop was developed to showcase the functionality and possible application of BCL::Cheminform to the community of the Vanderbilt Institute of Chemical Biology. An in-depth description of the workshop tutorial is provided in the Appendix. Another scenario could entail providing LB-CADD functionality as a service. Trained cheminformatics researchers who are familiar with the functionality of LB-CADD methods could reach out and apply novel LB-CADD tools collaboratively with medicinal chemists. This approach would maximize the success rate of applying all tools correctly while minimizing the frustration level of users when dealing with unfamiliar software.

In summary, the BCL::Cheminfo software suite provides an ideal framework for QSAR modeling and virtual high-throughput screening that is easily extendable to master the aforementioned challenges. The software is freely available for the academic community and will hopefully be deemed useful to the scientific community in future years to come.

APPENDIX

General Comments

Each section of the appendix is structured according to chapters. For each of the four main chapters a paragraph with the supplemental material is provided. Additional paragraphs summarize projects, ideas, or experiments that are not yet published or are incomplete. All computational work presented here is based on applications programmed in the BCL. All command lines presented in this appendix are functional as of BCL v3.1.0 SVN revision r4799. A current executable is available. The appendix is accompanied by a DVD. Each respective paragraph spells out a specific link to the directory layout. General information for each directory is provided in a respective README file on the DVD.

Supporting information for Chapter 2

Ligand-based virtual high-throughput screening benchmark and method development

DVD path: /ligand-based_virtual_hts_benchmark

Molecular descriptors numerically encode chemical structure

A total of 60 descriptor groups with 1,284 numerical descriptors were implemented in this study (see Table 13). The 60 categories contain scalar descriptors such as molecular weight, number of hydrogen bond donors, -acceptors, octanol / water partition coefficient, total charge, and topological polar surface area. Nine additional chemical properties were computed for every atom including atom identities, σ -, π -, and total charges, σ -, π -, and lone pair electronegativities, effective atom polarizabilities, and VC2003 atom charges [125]. Three encoding functions (2D auto-correlation, 3D auto-correlation, radial distribution function) are paired with each of the chemical properties to yield 27 fingerprints [77, 80]. In addition, each fingerprint is computed a second time applying van der Waals surface area as a weight factor.

Analyzing the overlap of optimal descriptor sets reveals extent of similarity

Descriptor selection identified optimized descriptor sets for each PubChem data set and machine learning algorithm pairing. The following experiment analyses the pair-wise overlap of optimized descriptor sets for each PubChem dataset using Tanimoto coefficients. The average overlap ranges from 0.01 to 0.31 when comparing different machine learning algorithms on one PubChem dataset. This is indicative of chosen descriptor sets having only few descriptors in common for a pair of machine learning algorithms. Individual off-diagonal values never exceed a Tanimoto coefficient of 0.47 suggesting that two optimized descriptor sets have generally less than half of their descriptor values in common for a particular SAID. Also, the size of the optimal descriptor set varies for every machine learning algorithm. The diagonal values show the recovered descriptors in comparison to the entire available 1284 descriptor values. Naively, one might expect a strong dependence on the target data set. Thus, the relation between chemical

Table 13: Overview of descriptors by category, description, and number of descriptor features.

#Group	Descriptor Category	Descriptor Name	#Features
	Scalar descriptors		
1		Molecular weight of compound	1
2		Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule	1
3		Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule	1
4		Topological polar surface area in Å ² of the molecule derived from polar 2D fragments	1
5		Octanol/water partition coefficient in log units	1
6		Total charge of the molecule	1
	Vector descriptors		
7	2D autocorrelation	Atom identity	11
8		Sigma charge	11
9		Pi charge	11
10		Total charge	11
11		Sigma electronegativity	11
12		Pi electronegativity	11
13		Lone pair electronegativity	11
14		Polarizability	11
15		VC/2003 partial atom charges	11
16	2D autocorrelation	Atom identity	11
17	weighted with	Sigma charge	11
18	van der Waals	Pi charge	11
19	surface area	Total charge	11

Table 13 continued

20		Sigma electronegativity	11
21		Pi electronegativity	11
22		Lone pair electronegativity	11
23		Polarizability	11
24		VC/2003 partial atom charges	11
25	3D autocorrelation	Atom identity	12
26		Sigma charge	12
27		Pi charge	12
28		Total charge	12
29		Sigma electronegativity	12
30		Pi electronegativity	12
31		Lone pair electronegativity	12
32		Polarizability	12
33		VC/2003 partial atom charges	12
34	3D autocorrelation	Atom identity	12
35	weighted with	Sigma charge	12
36	van der Waals	Pi charge	12
37	surface area	Total charge	12
38		Sigma electronegativity	12
39		Pi electronegativity	12
40		Lone pair electronegativity	12
41		Polarizability	12
42		VC/2003 partial atom charges	12
43	Radial Distribution	Atom identity	48
44	Function	Sigma charge	48
45		Pi charge	48
46		Total Charge	48
47		Sigma electronegativity	48
48		Pi electronegativity	48
49		Lone pair electronegativity	48
50		Polarizability	48
51		VC/2003 partial atom charges	48
52	Radial Distribution	Atom identity	48
53	Function	Sigma charge	48
54	weighted with	Pi charge	48
55	van der Waals	Total charge	48
56	surface area	Sigma electronegativity	48
57		Pi electronegativity	48
58		Lone pair electronegativity	48
59		Polarizability	48
60		VC/2003 partial atom charges	48
Total			1284

structure and biological activity can be established using multiple different combinations of descriptors – a finding that can be explained by linear dependence among different descriptor sets. Figure 17 compares descriptor set overlap from the perspective of each machine learning algorithm in the same fashion as in Table 13. Again, only few descriptors overlap. The extend of percent overlap compared to the initial 1284 descriptor values ranges from 0.3% to 83% (KN), 16% to 52% (ANN), 10% to 45% (DT), and 20% to 72% (SVM). ANN and DT tend to choose a more compact descriptor then KN. SVM shows a slightly increased descriptor set in comparison. High overlap does not seem to correlate with similar targets as seen for SAID 463087.

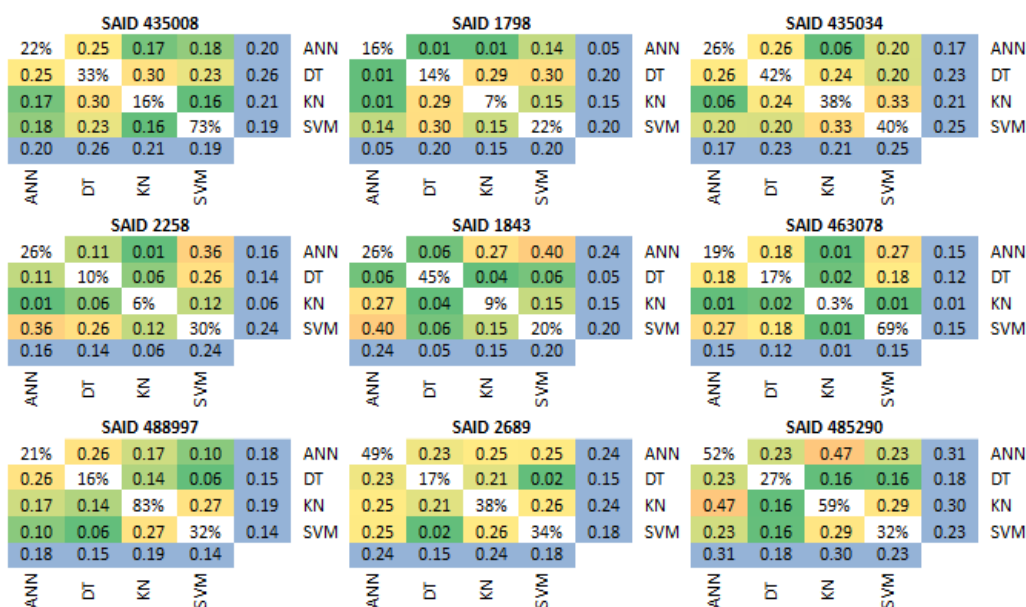


Figure 17: Pair-wise comparison of optimized descriptor sets.

Machine learning algorithms trained on the same PubChem data set were evaluated for descriptor set overlap. Each dataset (SAID) is represented with a heat map reporting the Tanimoto coefficients of element overlap for two machine learning methods shown in rainbow coloring. 0.0 (green) represents no overlap while 1.0 (red) indicates full overlap. The average percent overlap is depicted in blue taking off-diagonal values into account. All optimal descriptor sets determined by SFFS. The diagonal values (white) are excluded from the average calculation and represent the percent overlap of the optimized descriptor set compared to the initial set containing 1284 descriptors. This figure was adapted from [3].

Complete representation of inactive compounds improves QSAR model precision for initial true-positive range

Generally, experimental HTS datasets are highly unbalanced, i.e. the number of active compounds is much smaller than the number of inactive compounds. This poses a challenge for training predictors with naïve objective functions as – for example – 99% of all compounds are classified correctly by calling all compounds “inactive” and the root mean square difference (RMSD) between experimental and predicted activity will be low. To circumvent this problem all training data sets were balanced – i.e. an equal

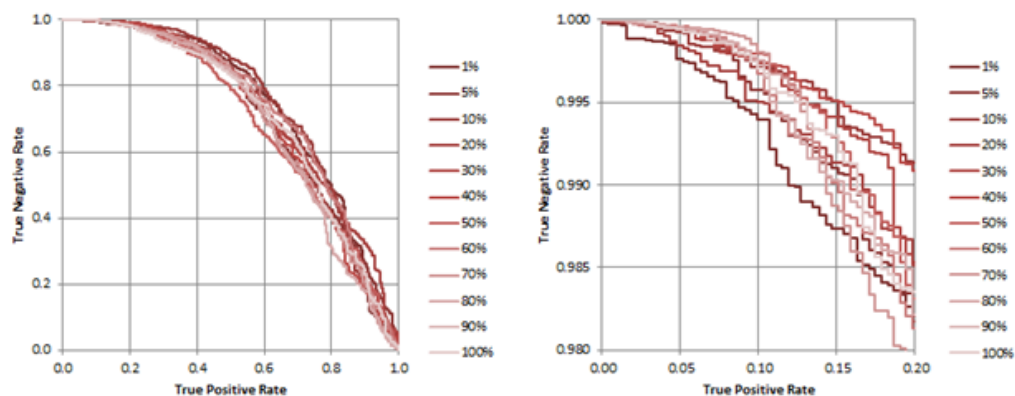


Figure 18: Heat maps reporting Tanimoto coefficients of descriptor overlap.

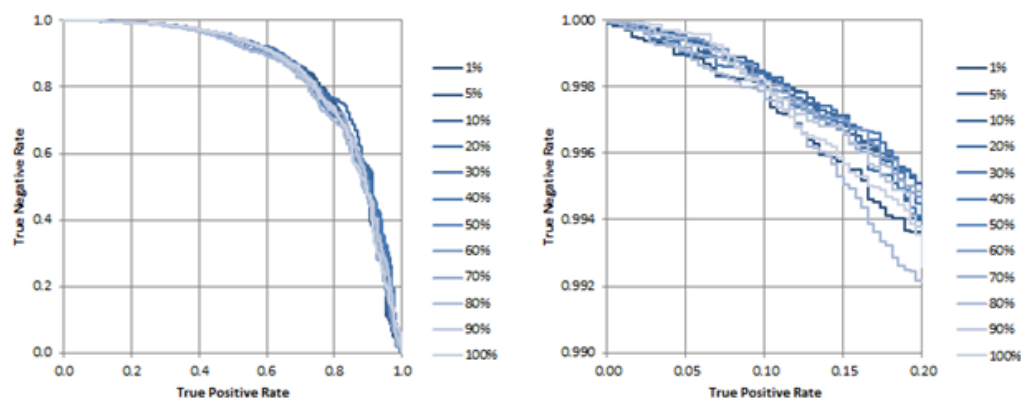
SFFS was applied to optimize descriptor sets among the various machine learning algorithms trained on different PubChem data sets (SAID). 0.0 (green) indicates no overlap while 1.0 (red) indicates full overlap. The average overlap is depicted in blue. All optimal descriptor sets were determined by SFFS. The diagonal values (white) are excluded from the average calculation and represent the percent of overlap of the optimized descriptor set compared to the initial set containing 1284 descriptors. This figure was adapted from [3].

number of active and inactive molecules were used. This can be achieved through over-sampling of active or under-sampling of inactive molecules. The first strategy utilizes all data for training while the second strategy only uses part of the inactive molecules, a strategy that could be attractive in order to accelerate the training procedure and create substantially smaller datasets. Note that the number of inactive compounds in the independent data set was not modified, i.e. the QSAR model still needs to classify all compounds of the independent dataset correctly. The percentage of inactive compounds reserved for training and monitoring data set partitions was systematically increased ranging from 1%, 5% and 10% to 100% in 10% steps for three selected datasets with SAIDs 488997, 485290 and 2258. Each data set contains more than 300,000 inactive compounds total. Adding more inactive compounds to the training and monitoring dataset increased the integral of the TNR-TPR curve for the initial TPR range compared to the 1% case (see Figure 19). A steeper TNR-TPR curve indicates a high TP to FP ratio ($FP=1-TN$), implies a high precision and therefore a high Enrichment. Evaluating the TPR range of 0.0 to 0.2, QSAR models with more inactives included in the training and monitoring data showed a higher precision for identifying active compounds compared to the 1% case. These results suggest a similar overall performance for all training data configurations but higher accuracy for the initial TPR range if more inactives are involved.

SAID 488997



SAID 485290



SAID 2258

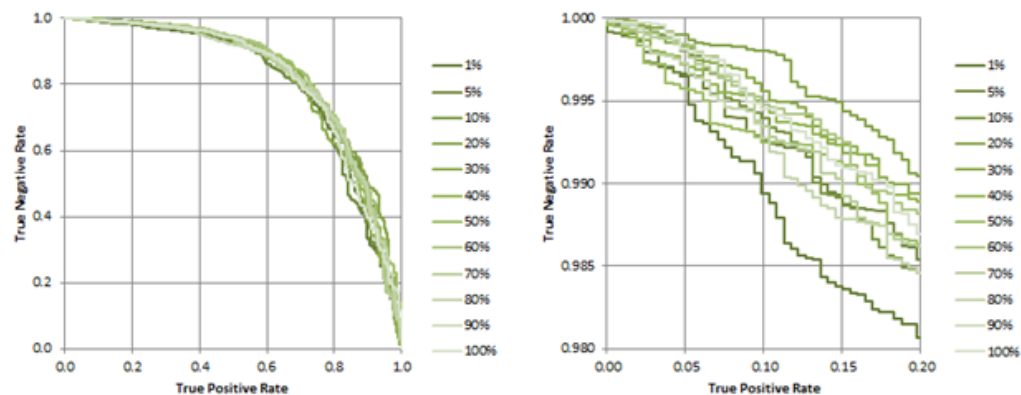


Figure 19: The TNR-TPR curve is shown for SAIDs 488997, 485290, and 2258.

Each curve represents a QSAR model trained on 1%, 5%, 10% ... 100% of the given monitoring and training data. The plots on the right-hand side show an enlarged view of the initial portion of the true-positive rate. With increasing percent of the used monitoring and training data the color intensity of each curve decreases. (1% - dark, 100% - light). This figure was adapted from [3].

Protocol Capture

The protocol capture was conducted on Linux CentOS 5. bcl version: 2.5.0; bcl svn revision: 4313. Every step assumes your root directory is the top level directory of the protocol capture directory!

The following steps capture the procedure for one of the benchmark data sets (eg. SAID 1798). All other benchmark data sets can be processed in the same way.

a) Data set generation: download AIDs

```
cd /bin/dataset_generation/
```

```
./aid_download.sh YOUR_AID_NUMBER
```

The script `aid_download.sh` will download the molecules associated with a specific PubChem AID. The result is a `.sdf.gz` file with all relevant molecules and a `.csv` file containing the biological data for every compound. `YOUR_AID_NUMBER` is the PubChem AID of interest.

b) Data set generation: process PubChem AIDs

```
cd /bin/dataset_generation/
```

```
./molecule_pipeline.sh YOUR_AID_NUMBER.csv YOUR_AID_NUMBER.sdf.gz
```

```
#the generate .bin files have to be copied to the /bin/cross_validation_pipeline/data directory
```

```
mv ./YOUR_AID_NUMBER_actives.bin ../cross_validation_pipeline/data
```

```
mv ./YOUR_AID_NUMBER_inactives.bin ../cross_validation_pipeline/data
```

`YOUR_AID_NUMBER` is the PubChem AID of interest. The script `molecule_pipeline.sh` will clean up all molecules by removing duplicates, generating 3D coordinates with CORINA, randomize the order and separate actives from inactives. If active compounds have no biological value (EC50/IC50) assigned, then a given value is set. binary files (`.bin`) containing small molecule descriptors and the associated biological

data will be generated for machine learning training .bin files have to be present in the /bin/cross_validation_pipeline/data directory.

c) Descriptor selection preparation

it is recommended to copy and rename the directory /bin/cross_validation_pipeline to represent the data set designation! The protocol capture will omit this step!

```
cd /bin/cross_validation_pipeline
```

make sure that the bcl.exe symbolic link is valid in bcl/ ; edit the file include.sh and set all variables accordingly; adjust the variable *dataid* to follow the pattern 'aidYOUR_AID_NUMBER', an example is given in the file for aid891; variable *dataset_size* should approximate the number of compounds used in both .bin files to request the right amount of memory from the pbs scheduler; the section TRAINING OBJ, CUTOFF and PARITY should set the variable cutoff and parity properly and only one of the objective function string should be unlocked! The training objective function will be applied in every iteration step during training; the section ITERATE specifies the chosen machine learning technique. only one set of the variables learning, iterate, and training_chunk_composition; the section FINAL OBJ FUNCTION specifies the objective function consisting of variables *obj_final* and *obj_final_prefix* applied only once at the end of ML training all remaining variables should be set accordingly to their description in *include.sh*

The set of scripts in /bin/cross_validation_pipeline makes it possible to perform descriptor selection (IG, FS, SFFS) and consensus predictions.

d) Descriptor selection

```
cd /bin/cross_validation_pipeline
```

get all options for descriptor selection by IG (information gain) and FS (fscore):

```
./submit_descriptor_reduction.sh
```

launch descriptor selection FS execute:

```
./submit_descriptor_reduction.sh fscore 2 local 10 600
```

launch descriptor selection IG execute:

```
./submit_descriptor_reduction.sh infogain 2 local 10 600
```

launch descriptor selection by SFFS (sequential feature forward selection) you need to have access to a pbs scheduler queue!

The type (SFFS) was set in the include.sh script!

```
./submit_descriptor_selection.sh start
```

Once the descriptor selection has stopped you can retrieve results:

```
cd results/
```

for IG and FS:

```
./results_descriptor_reduction.sh
```

for SFFS:

```
./results_descriptor_selection.sh
```

After choosing your descriptor selection method of choice, start your descriptor selection run.

e) Cross-validation

```
cd /bin/cross_validation_pipeline
```

to determine the best performing descriptor set

if you ran descriptor selection by IG or FS retrieve your final descriptor set with:

```
./get_best_descriptors_by_reduction.sh fscore
```

OR

```
./get_best_descriptors_by_reduction.sh infogain
```

if you ran descriptor selection by SFFS retrieve your final descriptor set with:

```
./get_best_descriptors.sh
```

to start cross-validation with the best performing descriptor set launch:

```
./submit_cross_validation.sh
```

to determine the results of the cross-validation run, execute:

```
cd results
```

```
./result_cross_validation.sh roc
```

results should contain a gnuplot script that can be executed to obtain a graphical representation (.png)

```
gnuplot cv_result.gz.gnuplot
```

f) Consensus Prediction

```
cd /bin/cross_validation_pipeline
```

to determine a consensus prediction, copy all available cv_results.gz to a separate directory (eg. /bin/consensus_prediction) and rename each filename to contain the machine learning technique.

```
./bcl.exe ComputejuryStatistics -input `ls /bin/consensus_prediction` -potency_cutoff 4.0 -table_name  
table.txt
```

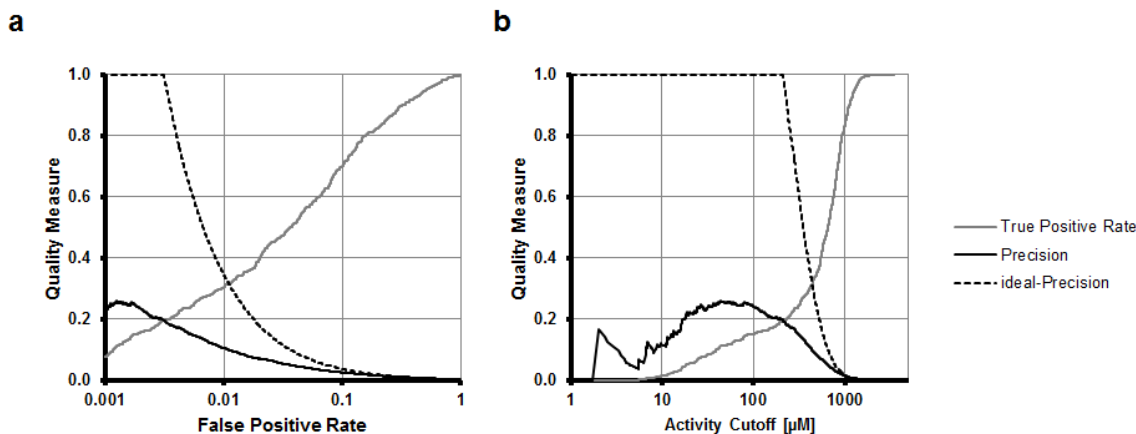


Figure 20: ROC curve evaluating the consensus model.

In Figure 20 a) the consensus QSAR model is quantified by a receiver operating characteristic (ROC) curve analysis. The plot illustrates the quality measures True Positive Rate (grey) and Precision (black) on a logarithmic scale. The ideal Precision plot (black dashed) is shown as a baseline. True Positive Rate ($= TP/(TP+FN)$) shows the ratio between the true positive active compounds and all positive compounds, true positives and false negatives. The higher the True Positive Rate the more accurate the QSAR model is in identifying true active compounds. Precision ($= TP/(TP+FP)$) provides the ratio of true positive active compounds and all positive predicted compounds, true positives and false positives. The quality measure Precision is equivalent to a normalized form of Enrichment ($= Precision/(P/(P+N))$). For the data set of 530 actives (P) / 144,300 inactives (N), a precision of 16% corresponds to an Enrichment value of 40 fold at a False Positive Rate ($= FP/(FP+TN)$) threshold of 1%. While misclassifying 1% of the active compounds the QSAR model identifies 16% of the active compounds correctly. In Figure 20 b) the same quality measures are evaluated with respect to the prediction activity cutoff in μM on a logarithmic scale. Both sub Figure 20 a) and b) plot the quality measures True Positive Rate, Precision, and Information Gain Ratio on a logarithmic scale. As a reference, the ideal trajectory of the precision curve is given for comparison to an ideal predictor. Figure 20 a) plots all quality measures in respect to the false positive rate assuming an activity cutoff of $100\mu\text{M}$. Figure 20 b) shows the same quality measures as a function of

the activity cutoff. This evaluation proves useful when choosing a concentration threshold for virtual screening. The application ComputeJuryStatistics takes all raw experimental/predicted values of available cross-validation runs and computes the consensus between all possible combinations or raw experimental/predicted value files.

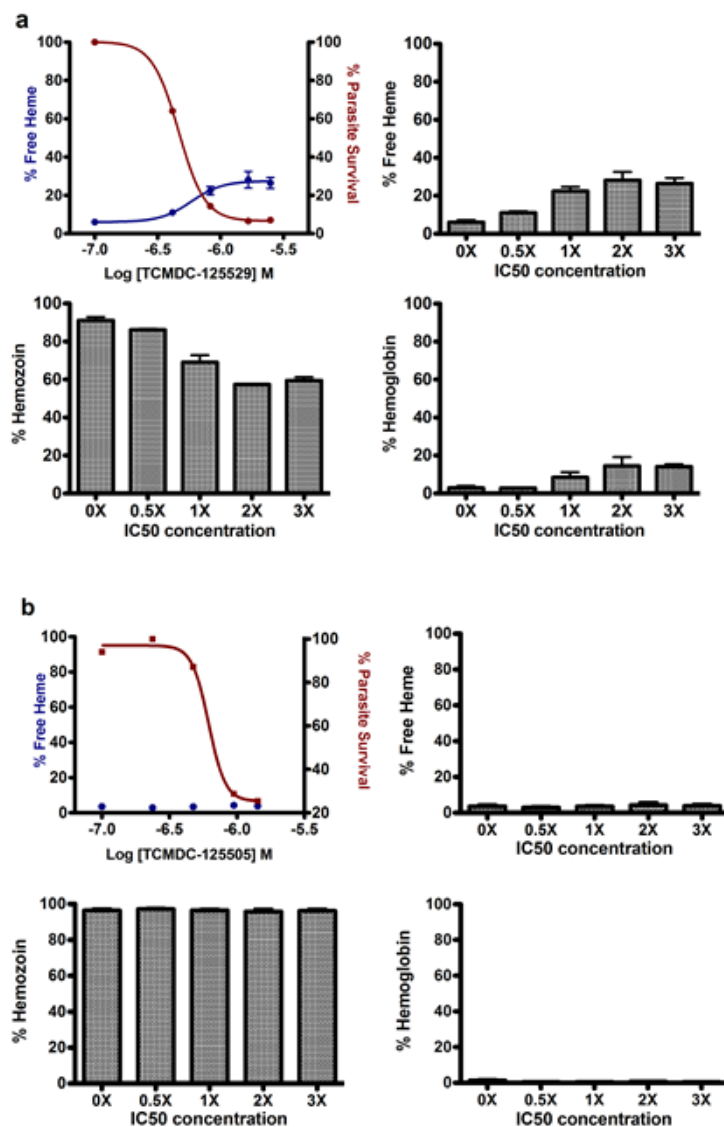


Figure 21: The heme speciation assay conducted on 17 β -hematin inhibiting GSK compounds.

Supporting information for Chapter 3

Identification of pathway specific inhibitors for β -hematin crystallization in plasmodium falciparum involved in Malaria

DVD path: /inhibitors_for_beta-hematin_crystallization_malaria

Results from a heme speciation assay

Figure 21 shows supplemental results from a heme speciation assay developed by Combrinck et al. [150] conducted on 17 β -hematin inhibiting GSK compounds. The parasite survival curve and fractions of free heme, hemoglobin, and hemozoin are shown for Figure 21 a) a compound confirmed as hemozoin inhibitor (TCMDC-125529) and Figure 21 b) one that does not inhibit hemozoin formation within the parasite (TCMDC-125505).

Protocol capture

The protocol capture was conducted on Linux CentOS 5. bcl version: 3.1.0; Every step assumes your root directory is the top level directory of the protocol capture directory!

a) Dataset preparation

```
./bin/generate_data.sh
```

the script will standardize the actives.sdf.gz and inactives.sdf.gz files using the BCL. Afterwards, datasets represented by .bin files are generated.

output: ./input/*.bin; files containing data sets for active and inactive molecules.

b) Descriptor scoring by information gain

edit the config/config.ini file to specify best performing descriptor set:

```
./bin/submit.py -t dataset_scoring --config-file config/config.ini
```

check calculated score distribution with:

```
gthumb score.infogain.out.png
```

calculates descriptor scores for every column. The .png file plots all ordered scores and visualizes the score distribution. output: score.infogain.out containing information gain scores for every feature column.

c) Feature selection

```
./bin/submit.py -t feature_selection --config-file config/config.ini
```

Edit your config.ini to specify the feature selection method to determine the best descriptor set.

d) Cross-validation

```
./bin/submit.py -t cross_validation --config-file config/config.ini
```

Edit your config.ini to specify the best descriptor set.

e) Consensus prediction

Repeat step 3 for all other machine learning approaches (svm, kohonen, dt). Copy the cross-validated independent dataset evaluations for each machine learning method into a new directory.

Create a new directory 'consensus':

```
mkdir consensus
```

Name each independent prediction output conveniently the same as the machine learning method (ann, svm, kohonen, dt).

```
cp results/{ann|svm|kohonen|dt}/independent0-?_monitoring0-?_number0.gz.txt  
consensus/{ann|svm|kohonen|dt}
```

Use the cross-validated machine learning models for a consensus prediction

```
bcl.exe model:ComputeStatistics -input ann svm kohonen dt -potency_cutoff 4
```


The generated output table shows all machine learning consensus combinations and their respective quality measures.

Supporting information for Chapter 4

Novel allosteric modulators for mGlu5 acting on distinct site related to CPPHA binding

A protocol capture will be provided on the DVD.

DVD path: /modulators_mglu5_related_to_cppha_binding

Protocol capture

The protocol capture was conducted on Linux CentOS 5. bcl version: 3.1.0; Every step assumes your root directory is the top level directory of the protocol capture directory!

a) Dataset preparation

```
./bin/generate_data.sh
```

the script will standardize the actives.sdf.gz and inactives.sdf.gz files using the BCL. Afterwards, datasets represented by .bin files are generated.

output: ./input/*.bin; files containing data sets for active and inactive molecules.

b) Descriptor scoring by information gain

edit the config/config.ini file to specify best performing descriptor set:

```
./bin/submit.py -t dataset_scoring --config-file config/config.ini
```

check calculated score distribution with:

```
gthumb score.infogain.out.png
```

calculates descriptor scores for every column. The .png file plots all ordered scores and visualizes the score distribution. output: score.infogain.out containing information gain scores for every feature column.

c) Feature selection

```
./bin/submit.py -t feature_selection --config-file config/config.ini
```

Edit your config.ini to specify the feature selection method to determine the best descriptor set.

d) Cross-validation

```
./bin/submit.py -t cross_validation --config-file config/config.ini
```

Edit your config.ini to specify the best descriptor set.

e) Consensus prediction

Repeat step 3 for all other machine learning approaches (svm, kohonen, dt). Copy the cross-validated independent dataset evaluations for each machine learning method into a new directory.

Create a new directory 'consensus':

```
mkdir consensus
```

Name each independent prediction output conveniently the same as the machine learning method (ann, svm, kohonen, dt).

```
cp results/{ann|svm|kohonen|dt}/independent0-?_monitoring0-?_number0.gz.txt  
consensus/{ann|svm|kohonen|dt}
```

Use the cross-validated machine learning models for a consensus prediction

```
bcl.exe model:ComputeStatistics -input ann svm kohonen dt -potency_cutoff 4
```

The generated output table shows all machine learning consensus combinations and their respective quality measures.

Supporting information for Chapter 5

Small molecules property predictions

A protocol capture will be provided on the DVD.

DVD path: /small_molecule_property_predictions

The same protocol was used to train QSPR model for both publications [1] and [2]. Protocol Capture

The protocol capture was conducted on Linux CentOS 5. bcl version: 2.5.0; bcl svn revision: 4313. Every step assumes your root directory is the top level directory of the protocol capture directory!

The raw data is given by two .sdf.gz files with all relevant molecules containing logP data. (MiscProperty: P_ow)

It is recommended to copy and rename the directory /bin/cross_validation_pipeline to represent the data set designation! The protocol capture will omit this step!

a) Set up scripts

```
cd /bin/cross_validation_pipeline
```

Variable dataset_size should approximate the number of compounds used in both .bin files to request the right amount of memory from the pbs scheduler

The section TRAINING OBJ, CUTOFF AND PARITY should set the variable cutoff and parity properly and only one of the objective function strings should be unlocked! The training objective function will be applied in every iteration step during training. The section ITERATE specifies the chosen machine learning technique. only one set of the variables learning, iterate, and training_chunk_composition. The section FINAL OBJ FUNCTION specifies the objective function consisting of variables obj_final and

obj_final_prefix applied only once at the end of ML training. All remaining variables should be set accordingly to their description in include.sh. The set of scripts in /bin/cross_validation_pipeline makes it possible to perform descriptor selection (SFFS) and consensus predictions. The include.sh script is the main configuration file.

b) Cross-validation

Start cross-validation with the best performing descriptor set launch:

```
./submit_cross_validation.sh
```

Determine the results of the cross-validation run, execute:

```
cd results
```

```
./result_cross_validation.sh roc
```

The results should contain a gnuplot script that can be executed to obtain a graphical representation (.png)

```
gnuplot cv_result.gz.gnuplot
```

Once the final descriptor set is determined a full 10x9 cross-validation can be applied to determine the objective function of the final cross-validated model. The cross-validated models based on the best performing descriptor set will be stored in the MySQL database. After retrieving the results from the final cross-validation run a graphical representation can be generated through gnuplot.

c) Consensus prediction

```
cd /bin/cross_validation_pipeline
```

To determine a consensus prediction, copy all available cv_results.gz to a separate directory (eg. /bin/consensus_prediction) and rename each filename to contain the machine learning technique.

```
./bcl.exe ComputejuryStatistics -input `ls /bin/consensus_prediction` -potency_cutoff 4.0 -table_name  
table.txt
```

The application ComputejuryStatistics takes all raw experimental/predicted values of available cross-validation runs and computes the consensus between all possible combinations or raw experimental/predicted value files.

Tutorial for QSAR modeling, virtual screening, and cluster analysis with BCL::ChemInfo

This tutorial is part of a workshop to disseminate the functionality of BCL::ChemInfo to the Vanderbilt community.

DVD path: /vHTS_workshop

The tutorial starts with a file that contains the results of an HTS screen: 9,122 molecules are classified as active or inactive, according to their activity in the screen. We train a QSAR model to recognize molecules according to their activity, and then apply that model to virtually screen a larger library of 100,000 molecules.

Hands-on 1: Data set preparation

In this section, you will prepare a data set for QSAR model construction starting from an SDF file format.

The files required for this lesson are found in the sub directory *qsar_tutorial*.

Preparing the SDF file

The SDF file named **initial_hts_result.sdf** contains connectivity information describing the molecules from an HTS screen as well as the resulting experimental bioassay data. If you wish, you can open this text file in a text editor to examine the contents. The experimental data is included as an SDF miscellaneous property string that looks like this: “> <Activity>”. Some compounds may be missing this property. The first step is to filter the SDF file to ensure that all molecules can be properly interpreted by

the BCL, by removing any compounds that are missing the Activity property. This is performed using the following command line:

```
bcl.exe molecule:Filter -input_filenames initial_hts_result.sdf -has_properties 'Activity' -  
output_matched molecules_with_activity_property.sdf -output_unmatched  
molecules_missing_activity_property.sdf
```

This will divide your compounds into two files. Were any molecules discarded during this process? If so, they can be found in this file:

```
molecules_missing_activity_property.sdf
```

You can count the number of records in that file by grepping for the record separator:

```
grep -c '$$$$' molecules_missing_activity_property.sdf
```

Execute the following command to find about about the activity distribution of the compounds:

```
bcl.exe molecule:Properties -input_filenames molecules_with_activity_property.sdf -  
numeric_histogram 'Activity' 0 10 20 -output_histogram
```

Determine the cutoff value by examining the histogram file. What is a suitable cutoff value?

Next, two SDF files will be created to separate the “actives” from the “inactives” depending on your cutoff value \$CUTOFF that you determined from examining the histogram (we chose 45 micromolar):

```
bcl.exe molecule:Filter -input_filenames molecules_with_activity_property.sdf -output_matched  
molecules_actives.sdf -output_unmatched molecules_inactives.sdf -compare_property_values  
'Activity' less 45
```

Now check-the number of molecules in each of the created actives.sdf and inactives.sdf files using “grep” as above. How many molecules are contained in each file?

```
grep -c '$$$$' molecules_{in}actives.sdf
```

Perform clustering analysis

Now that we have ensured that the molecules are usable by the BCL and have been separated into actives and inactives, let's visually assess the chemical space of the actives through hierarchical agglomerative clustering analysis. Run the following script to create the dendrogram for visualization:

Copy your molecules_with_activity.sdf file into the directory called **clustering**:

```
cp molecules_actives.sdf clustering/
```

Change into directory clustering.

```
cd clustering
```

Start processing your molecules by executing a molecule clean up script:

```
clean_up_molecules.sh molecules_actives.sdf
```

This will create an output file (and it automatically zips the input and output files for you):

```
molecules_actives_clean.sdf.gz
```

Start clustering your cleaned up molecules by executing this command to learn about all the options available for the clustering command.:

```
cluster_molecules.py -h
```

The following clustering run should take about 20 minutes, using a linkage value of 0.3, a sampling factor of 0.1 (which means it uses 10% of the data to construct fragments to determine common scaffolds), and uses a cluster size cutoff of 5. These values may be adjusted for different datasets. Execute following command:

```
cluster_molecules.py -m molecules_actives_clean.sdf.gz -l 0.3 -s 0.1 -c 5
```

You will see the program 1) build a fragment library, 2) do several filtering steps 3) construct the distance matrix 4) perform the actual clustering step. The output to the screen indicates the number of clusters, and the number of compounds per cluster (Size). You can adjust the parameters and re-run the clustering if you get an unreasonably small or large number of clusters.

To identify the main scaffold in each cluster issue the following command:

```
cluster_scaffolds.sh
```

A new directory will be created containing all clusters and cluster scaffolds as sdf files

```
cd cluster_sdf
```

Now issue the following command to visualize the dendrogram of the scaffolds:

```
pymol dendrogram.py
```

Look at the individual scaffold molecules with:

```
pymol scaffolds.sdf
```

You can step through the scaffolds using the movie commands in the bottom-right corner of the PyMOL window. Note that the scaffolds do not have Hydrogens added.

It is important that the actives chemical space is diverse to ensure generalizability of the model. If all of the scaffolds are similar, it will be difficult to apply the QSAR model to external databases successfully to find novel scaffolds.

After visualizing the diversity of the scaffolds, you can go back to the main tutorial directory (qsar_workshop) with this command:

```
cd ../../
```

Descriptor generation

Numerical descriptors of the molecules will be calculated to serve as input for the QSAR method. The descriptors to be calculated are shown below

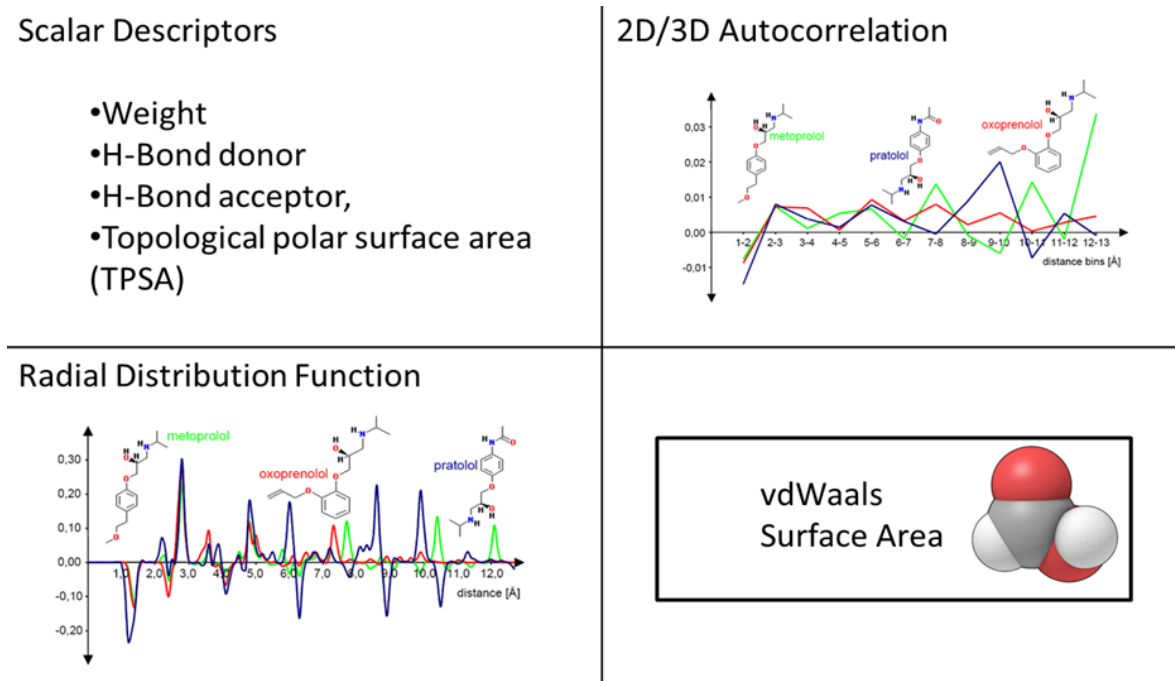


Figure 22: Graphical overview of molecular descriptors applied in the tutorial.

Now, copy our active and inactive files into our next working directory:

```
cp molecules_actives.sdf qsar_cheminfo/data
cp molecules_inactives.sdf qsar_cheminfo/data
cd qsar_cheminfo
```

Have a look into the files *features.object* and *results.object* to find out about the descriptors used in this dataset. This command will display the list of molecular descriptors that we will include in the model:

```
less features.object
```

And this command will show the transformation of Activity values from the original micromolar scale to a log scale that is convenient for training the model:

```
less results.object
```

Run the following two commands to generate two dataset files, "the binfiles." Those files contain the scaled activities, and values for each descriptor calculated from the structures in your input sdf files (of your actives and inactives). This step should take just a few minutes on our small file of actives, and ten times as long for the larger inactives file:

```
bcl.exe descriptor:GenerateDataset -source 'SdfFile(filename=data/molecules_actives.sdf)' -  
feature_labels features.object -result_labels results.object -output data/molecules_actives.bin -  
scheduler PThread 2
```

and then:

```
bcl.exe descriptor:GenerateDataset -source 'SdfFile(filename=data/molecules_inactives.sdf)' -  
feature_labels features.object -result_labels results.object -output data/molecules_inactives.bin -  
scheduler PThread 2
```

If you watch the output from this command, you will see that a few molecules may get rejected from the dataset generation. This happens when the molecule contains unrecognized atomtypes. The model will still be robust even if a few molecules are discarded. The resulting output file is a binary format and is not human readable. However, the descriptors can be viewed using the following command:

```
bcl.exe descriptor:GenerateDataset -compare data/molecules_actives.bin | less -S  
and:
```

```
bcl.exe descriptor:GenerateDataset -compare data/molecules_inactives.bin | less -S
```

Scroll down to the section that starts 'Statistics.' The following is a list of molecular descriptors, which you will recognize from the lecture. You can also scroll to the right to see the column headings: "Ave Std Min Max" which indicate the range of values for each descriptor. The last line is the range of scaled Activity values in your dataset.

Hands-on 2: QSAR model construction

Now that the data set of molecular descriptors has been created, a QSAR model must be constructed. We will now construct a QSAR model using an Artificial Neural Network (ANN).

Training your first model

You will now use the model::Train bcl application to train an ANN for distinguishing actives from inactives. First, type:

```
bcl.exe model:Train -help | less
```

to look at the possible arguments for this application. As the help output indicates, in many cases you can get help in greater detail for each individual argument. To train an ANN using enrichment as an objective function we will use a submission script that does the heavy lifting for us.

As you saw when looking at the help function from BCL model:Train, there are a number of options available. To make it more manageable, we will use a subset of those options in a configuration file. Edit the file config.ini using gedit or vim, and make sure the following options are set:

```
[datasets]
datasets: ['data/molecules_actives.bin','data/molecules_inactives.bin']

[cv]
cross-validations: 10
monitoring-id-range: [ 0, 0]
independent-id-range: [ 1, 1]
id: qsar_workshop
```

Note that here we are telling BCL to divide the dataset into ten equal-sized "chunks," the first of which is used for monitoring, and the second for independent testing. Then execute training of the model with all possible features:

```
./submit.py -t cross_validation
```

The command above runs a script to perform 10-fold cross validation (discussed later) and uses 80% training, 10% monitoring, and 10% independent as was discussed in the previous lecture. This step should take just a few minutes, and will create directories named results/ log_files/ and models/.

Analyze full model results

To analyze the performance of the model, we can create a receiver operator characteristic (ROC) curve. This plots the true positive rate versus the false positive rate. Run the following commands to create the ROC curve for visualizing the QSAR model performance.

Go into your results directory:

```
cd results/cv_qsar_workshop
```

and examine the roc curve in the created .png file (eog is a png viewer):

```
eog independent1-1_monitoring0-0.gz.txt.gnuplot_txt.png
```

The red curve is your ROC curve (False Positive Rate vs. True Positive Rate), the green curve is the normalized enrichment value (precision) for that particular FPR cutoff, and can be compared to the blue ideal curve.

Hands-on 3: Feature selection

We have now created a baseline QSAR model using an ANN. The next step is to optimize the performance of this model through feature selection. The feature selection process to be used in this

section utilizes information theory to determine the relevance of each descriptor and eliminate those which contribute less to describing the observed activity.

Create feature scores

We first calculate information gain scores for each feature. Go back up to the `qsar_cheminfo` directory:

```
cd ../../
```

edit the file **config.ini** and make sure the following options are set as listed:

```
[score]
scoring-type: InformationGain
output_score_file: score.infogain.out
features: features.object
```

To calculate the information gain score execute following command:

```
submit.py -t dataset_scoring
```

A scores file called **score.infogain.out** will be generated. You can examine the file with the following command:

```
less -S score.infogain.out
```

Analyze feature scores

We can see the amount of discrimination that each feature (descriptor) contributes to the model by looking at the sorted values in the resulting graph:

```
display score.infogain.out.png
```

Each point on this graph represents an individual descriptor column in the dataset. The Y-value shows the amount of information gain from that descriptor. From this plot, you can estimate how many descriptors you need to perform good discrimination. This will be done in the next step.

Feature selection

We now use the calculated scores to perform a feature selection protocol. Now, three hundred QSAR models will be trained using different sets of features chosen by information gain score to optimize model performance.

edit the file **config.ini** and make sure the following options are set as listed:

```
[feature-selection]
range: [150,1284,150]

[cv]
monitoring-id-range: [0,4]
independent-id-range: [0,4]
cross-validations: 5
store-model: File
id: qsar_feature_selection
```

Executing the following script will perform the feature selection:

```
./submit.py -t feature_selection
```

This feature selection protocol will take ~1hr, so let's break for lunch at this time.

After this process completes, look in the results directory, at the final result file:

```
less -S results/final_cv_qsar_workshop_InformationGain_result.txt
```

The first field of the last line specifies the directory containing the best-performing descriptor set. Make note of the number. You will use that in the next step, where we call it `cv_qsar_workshop_InformationGain_top_???_features`. (???-should be substituted by best round number)

Hands-on 4: Cross-validation

Now that model optimization has been achieved through feature selection, we must extensively cross validate the QSAR model using the optimized feature set.

Cross validate optimized models

To perform a full cross validation, the script below will train 90 models. This resulting ensemble of optimized models maximizes generalizability by using the arithmetic mean for prediction. In order to use the optimized feature set, we need to copy that result file over:

```
cp models/cv_qsar_workshop_InformationGain_top_???.features/model000001.descriptor
best.descriptor
```

edit the file **config.ini** and make sure the following options are set as listed:

```
[score]
scoring-type: InformationGain
output_score_file: score.infogain.out
features: best.descriptor
```

```
[cv]
monitoring-id-range: [0,9]
independent-id-range: [0,9]
cross-validations: 10
store-model: File
id: qsar_final
```

Executing the following script will perform the feature selection:

```
./submit.py -t cross_validation
```

This step should take around an hour.

Analyze optimized results vs. full model results

Now that the feature selection and full cross validation has completed, we can create the ROC curve for the optimized model.

ROC curves have now been generated by the cross-validation script. Look in the ./results directory and the appropriate sub directory for .png files. How does this ROC curve compare with that achieved previously using the entire feature set?

Hands-on 5: Virtual screening of external libraries

Now that we've constructed optimized, cross-validated QSAR models using decision trees, we can use these models to predict active compounds for this target using large external libraries as input.

Run the vHTS using optimized models

Using the "external_lib.sdf" SDF file run the following command:

```
bcl.exe molecule:Properties -input_filenames external_lib.sdf -add  
'Mean(PredictedActivity(storage=File(directory=models/cv_qsar_final,prefix=model)))' -rename  
'Mean(PredictedActivity(storage=File(directory=models/cv_qsar_final,prefix=model)))'  
pred_activity_rank -output vhts.sdf.gz -input_start 1 -input_max 2000
```

This command will generate the optimized descriptors for the molecules in the SDF file, predict the activity using the 90-model ensemble, and insert the activities into the compressed SDF file "vhts.sdf.gz".

This process should take 30 minutes. Note that we are abbreviating this screen- we limit ourselves to the first 2000 compounds, so that we can finish in a short time for the workshop.

Clustering analysis of top actives

Now let's analyze the molecules predicted to be active. First, filter the predicted activities to a list of the top 1000 actives:

```
bcl.exe molecule:Reorder -input_filenames vhts.sdf.gz -sort 'MiscProperty("pred_activity_rank")' -  
reverse -output vhts.sorted.sdf.gz
```


Extract the top 1000 molecules:

```
bcl.exe molecule:Properties -add Index -input_filenames vhts.sorted.sdf.gz -input_max 1000 -  
output vhts.top1000.sdf.gz
```

Now, perform a clustering analysis by moving the files into the right directory:

```
cp vhts.top1000.sdf.gz ../cluster_top_vhts_hits
```

then go to the clustering directory:

```
cd ../cluster_top_vhts_hits
```

and execute the following commands:

```
./clean_up_molecules.sh vhts.top1000.sdf.gz  
cluster_molecules.py -m vhts.top1000_clean.sdf.gz -l 0.3 -s 0.1 -c 5  
cluster_scaffolds.sh
```

Visualize the resulting dendrogram with the retrieved scaffolds:

```
cd cluster_sdf/  
pymol dendrogram.py
```

Look at the individual scaffold molecules with:

```
pymol scaffolds.sdf
```

The individual molecules associated to each cluster can be found in the files

```
cluster_output_*.sdf and cluster_output_*.scaff.sdf
```

```
pymol cluster_output_<cluster_number>.sdf
```

and

cluster_output_<cluster_number>.scaff.sdf

You would then select the desired scaffolds and associated compounds for purchase.

Implementation of multi-output Support Vector Regression

This project describes the implementation of multi output support vector regression (SVR) in the BCL. A summer student in the Meiler laboratory, Georg Krause, worked under the supervision of Mariusz Butkiewicz on this project and provided the write-up on the DVD. Multi-output regression aims at learning a mapping from an input feature space to a multivariate output space. The output space is considered a sub manifold which incorporates its geometric structure into the regression process. A novel technique termed locally linear transformation (LLT) is used to specify a loss function on the output manifold. This method uses possible correlations between the SVR outputs and is described in [231]. The resulting convex optimization problem is solved by a sequential minimal optimization algorithm. This method is described for the single output SVR in [232]. The derivation of this algorithm for the method in [231] will be described here.

The derivation can be found in the MS word document on the DVD. See DVD path: /multi_output_svr

Implementation

Solving the sub problem

In our implementation we decided to set ϵ to 0. This simplifies the computation a lot, because then the whole sign dependent term vanishes. We first compute the optimal solution for $\vec{\alpha}_i$ and clip it to the bounds. Second we use the summation constraint to compute $\vec{\alpha}_j$ and check whether its boundaries are violated. In case of a violation we clip $\vec{\alpha}_j$ and adjust $\vec{\alpha}_i$ using the summation constraint again. Next we check for violations in $\vec{\alpha}_i$ and clip it if necessary. The checking is done until both alphas fit into boundaries within five iterations total at the maximum. If there is still a boundary violation, we ignore the summation constraint and clip both alphas independent from each other. At last we update all gradients to the new alphas using the following equation:

$$\Delta \vec{\varphi}_l = -\pi_i \Delta \vec{\alpha}_i k(x_i, x_l) - \pi_j \Delta \vec{\alpha}_j k(x_j, x_l).$$

Picking the two feature vectors i and j

For picking the two feature vectors to optimize I use a slightly modified method from [232]. i is picked so that $(\pi_i^T * (\varphi_i - b))^2$ is maximized ignoring the elements where the corresponding alpha element already hits its lower(if the current element of $\pi_i^T * (\varphi_i - b) < 0$)/upper bound. If we chose an i with a large error it is more likely to get a large error difference corresponding to a large change in alpha. We subtract the current bias from the error vector, because the alphas are not to optimize in a way they compensate the bias. The j is picked, so that $(\Delta\beta)^2 = (\pi_i \Delta\alpha_i)^2$ is maximized, yielding a maximal reduction of the error vectors. Therefore we solve the problem for each possible j as described above. Then we apply the solution with the j yielding maximal $(\Delta\beta)^2$. This means we update the gradients and alphas.

Get the final support vector model

x_i is called a support vector if and only if

$$\vec{\beta}_i = \pi_i * \vec{\alpha}_i \neq 0.$$

To reduce the number of support vectors, we set the alpha elements with an absolute value smaller than a threshold epsilon to zero. This could be a more simple way than the epsilon term in the optimal solution. The prediction for a new input vector x is

$$f(x) = \vec{b} + \sum_{i=1}^K \vec{\beta}_i k(x_i, x)$$

So we need only the betas to make a prediction.

Explanation of variables and parameters

In the implementation there are certain parameters used. Let us have a look at their influence here. In Figure 23 below you see some training data points fitted by a graph. Some of them are support vectors, so

they contribute to the graph. The contribution of each support vector is shown by the dotted plots. These are kernel functions; here we used the RBF-Kernel. The amount of influence is determined by alpha. If this is smaller than a threshold ϵ the influence is ignored in the end. The maximal possible influence is C or $-C$ respectively. This is to avoid wrong predictions due to outliers. The bias b is the average error and therefore close to the average output and contributes as an offset. The alphas are not able and not made to take care of this offset. The last parameter shown in the figure gamma is specific to the RBF function and determines the range of influence. This should not go over the entire data set. The parameter Kappa is the number of nearest neighbors used to determine the coordinate systems. Set it to 1 if you want these coordinate systems to be identity matrices. For details about the coordinate system computation see [231].

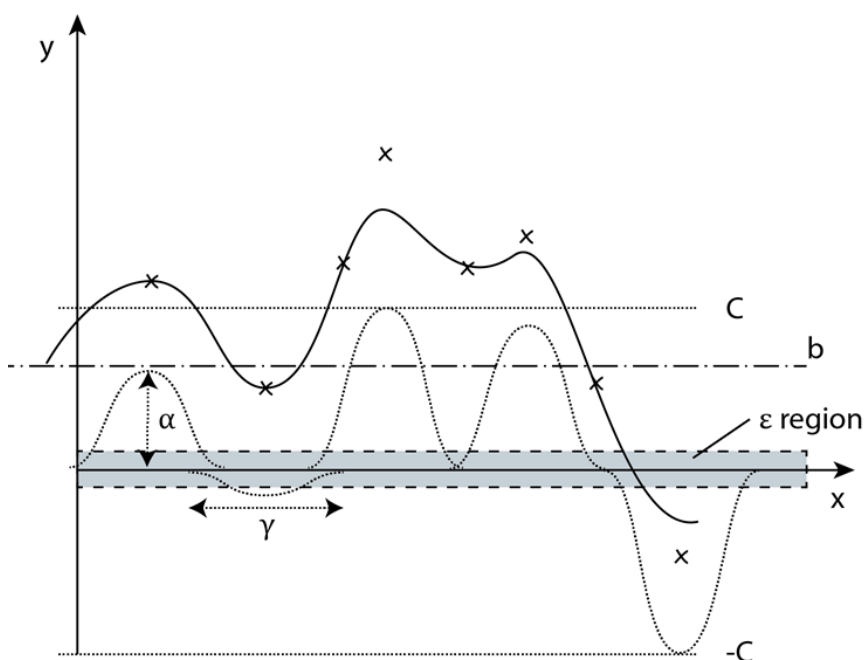


Figure 23: Explanatory schematic of variables used in multi-output SVR training.

Implementation of an auto-encoder algorithm for molecular descriptor compression

The auto-encoder algorithm is implemented in the BCL. An example setup can be found on the appendix DVD.

DVD path: /autoencoder

Deep learning allows for novel molecule encoding scheme to overcome the limitations of budget-constrained HTS campaigns.

QSAR modeling relies on the availability of HTS data representing the chemical space of the biological target [116]. The predictive accuracy of conventional QSAR modeling algorithms typically suffers from limited data of smaller HTS campaigns [116].

To address the bottleneck, we explore a novel strategy to build deep neural networks, based on stacking layers of denoising autoencoders (SdA). Deep learning refers to the training of multiple layers of hidden units in one network structure (see Aim Ia). SdAs are trained on corrupted versions of their inputs (noise) with the goal to recover the noise-free input from the trained hidden layer (denoising). The resulting algorithm is a variation of the stacking of autoencoders. It has been shown on a benchmark of classification problems to yield significantly lower classification error [233]. Higher level data representations learned in this purely unsupervised fashion help boost the performance of subsequent machine learning predictors. Qualitative experiments show that in contrast to ordinary autoencoders, denoising autoencoders are able to learn edge detectors, which are derived from natural image patches and enable hierarchical extraction of image determinants. Lastly, SdA abilities also allow for the inclusion of larger stroke detectors from digit images, a different form of image determinants [233-235]. This work clearly establishes the value of using a denoising criterion as a tractable unsupervised objective to guide the learning of higher level data representations.

The hypothesis of this project is derived from the field of image recognition. Each layer of SdAs will be able to capture distinct aspects of the structural information available from drug-like small organic ligands in publically accessible databases. It is an unsupervised technique which relies solely on the structural information of small molecule ligands and thus is independent of any HTS campaign.

Molecule encoding is an alternative to conventional descriptor selection schemes for QSAR.

QSAR modeling requires molecular descriptors to train a machine learning model. With the development and availability of new molecular descriptors, the encoding of a small molecule ligand can get complex and time-consuming. Current descriptor selection algorithms evaluate combinations of molecular descriptors in an iterative fashion. In addition, this process is dependent on the current protein target and has to be repeated for every new biological system. With the proposed encoding scheme, only one master network of SdA layers has to be trained. This scheme is independent of HTS campaigns and can be applied to small molecule structures of any biological system. This master SdA network is able to determine the relevant small molecule features and therefore circumvents the time expensive descriptor selection process. Subsequently, a cross-validated QSAR model can be trained based on encoded molecules of the SdA network.

Molecule encoding scheme using Stacked denoising Autoencoder (SdA)

With the recent advent of deep learning it is possible to develop more sophisticated network architectures suitable for supervised and unsupervised learning. The foundation of the molecule encoding scheme builds upon a deep learning technology called Stacked denoising Autoencoder. SdAs employ hierarchical extraction of essential determinants from existing molecular descriptors. Each layer in the hierarchy can be stacked upon each other. The training takes place in a separate step for each SdA until its input layer resembles the output layer (see Figure 24). The addition of noise to each input layer ensures a generalization ability of the SdA to prevent overfitting (memorizing). This pre-training procedure of each SdA is an unsupervised learning process which means no HTS association to a biological target is necessary. Only the molecular structure and properties are used for training each SdA layer. Thus, it is possible to leverage different aspects of chemical space, obtained from diverse and large public databases of small molecules, with each layer through dimensionality reduction. Subsequently, ligands evaluated through biological assay screens will be encoded using the proposed encoding scheme to develop

quantitative structure activity relationship (QSAR) models. These models correlate chemical structure (encoded ligand) with its biological activity for a specific biological target (protein).

The development of the proposed molecule encoding scheme requires access to small organic ligand structures representing the largest possible chemical space. Two of the largest publically available databases, Pubchem [236] and Zinc [237] will be applied. Each database contains characterized chemical compounds, 31 and 21 million compounds respectively.

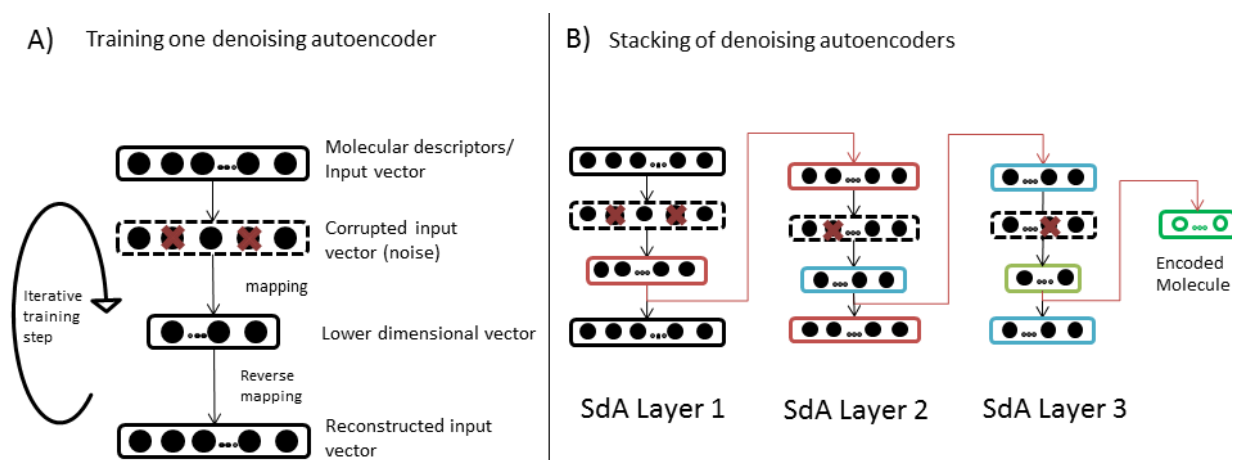


Figure 24: Schematic view of training de-noising autoencoders.

A) Flowchart showcasing the training of one denoising autoencoder. Molecular descriptors are altered to introduce a noise. The corrupted input is mapped into a lower dimensional hidden layer. A reverse mapping process seeks to reconstruct the original input vector from the lower dimensional hidden layer. B) Flowchart depicts the stacking aspect of training multiple layers of SdAs. Once a SdA layer is trained, its hidden layer output serves as input vector to the subsequent layer. The final layer encodes the molecule.

Each available compound will be represented by molecular descriptors expressing every molecule as a unique numerical fingerprint. Currently, the BCL provides implementations of 784 descriptor groups resulting in over 7,300 descriptor values.

Mapping of protein-ligand interactions elucidates determinants of structure-activity relations

The foundation of protein-ligand interaction mapping builds upon quantitative structure activity relationships (QSARs) that correlate chemical structure (ligand) with its biological activity for a specific

biological target (receptor) based on machine learning algorithms. Furthermore, amino acid sequence descriptors for receptor variants and descriptors for ligand chemotypes will be concatenated into one numerical fingerprint. So far, traditional ligand-based virtual screening methods encode solely ligand information screened against a single biological target. Screening the same set of ligands against mutated representations of the same receptor target will disclose specific biological activities for each compound. This will allow for determination of ligand specific binding site interactions and adds a new dimension to conventional ligand based virtual screening.

Protein sequence and amino acid information of the receptor mutant are associated with each available ligand. The resulting descriptor fingerprint serves as input to machine learning algorithms that establish a QSAR model through the proposed consensus prediction scheme.

The hypothesis states that trained models will allow for correlation of ligand derivatization with protein mutation. A proof of concept experiment involves two distinct amino acid sequence mutations through site-directed mutagenesis of residues affecting ligand binding and receptor function. Ligand structures with known activity against one or both mutations have to be available. Now, a consensus QSAR model can be trained to distinguish ligands that bind at binding sites with one variation or both. A possible starting point is a similar experiment performed in the Conn laboratory at Vanderbilt. Double-point mutations of the metabotropic Glutamate Receptor subtype 5 were carried out to discover negative allosteric modulators (NAMs).

Further, the trained consensus QSAR models can now be applied to virtually screen external commercial databases (e.g. ChemBridge, ChemDiv) for potential hit compounds involving a potential binding site. A systematic moiety substitution of a ligand scaffold is also possible. All putative hits from this virtual screen can be confirmed at the Vanderbilt HTS center through dose-response experiments against all available receptor mutants. These mutation assays provide evidence of significant functional groups involved in ligand-protein binding.

An opposite application is to guide and prioritize experimental sequence mutations by virtual screening for active compounds in the presence of a specific amino acid change. The impact of the mutation towards the activity of ligands can be assessed and low impact binding mutations can be avoided. Both application strategies highlight the potential of the proposed protein-ligand interaction map.

A protein-ligand interaction map seeks to pin-point interactions between ligand and receptor

Every ligand – mutant receptor combination is represented as a data vector of which the amount is equal to the number of receptor mutations multiplied with the number of ligands. This setup ensures the activity association of all ligands toward all available mutations. Given a predefined ligand scaffold it is possible to screen modified derivatives for receptor binding activity. Changes in activity can now be related to the modification of the ligand in conjunction with the receptor mutation. Conclusions about moieties in the small molecule can be drawn that are crucial to specific residue binding.

On the other hand, the systematic mutation of residues can be evaluated through virtual screening against the given ligand library as well. Therefore the impact of specific residues for ligand binding can be assessed. This allows the development of a correlations map to highlight interactions from the perspective of a ligand structure as well as the receptor mutation. A possible mapping scheme is shown in Figure 25 showing modification suggestions as a map on the ligand scaffold related to the influence of mutated residues.

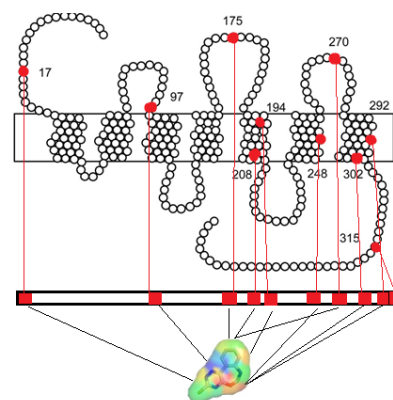


Figure 25: Schematic view of protein-ligand interaction. Amino acid mutations impact is correlated to ligand moiety influence shown as a pharmacophore map.

The resulting receptor – ligand map is expected to guide prospective experimental mutation studies to assay structural determinants for ligand binding. In addition to the systematic moiety substitution, the

trained consensus QSAR models can now be applied to virtually screen external commercial databases (e.g. ChemBridge, ChemDiv) for potential hit compounds involving a particular binding site.

Multi-output prediction QSAR model to create a small molecule prediction profile for mGlu₁₋₈

The goal of this project is to create a QSAR model to predict a small molecule activity profile involving mGlu receptor subtypes 1-8. The ChEMBL database contains molecule information in publications reporting compounds with activity against various mGlu Receptors. A total of 1,008 molecules was identified that showed at least activity against one mGlu receptor subtype.

A description to train a preliminary QSAR model and the protocol to acquire all 1,008 compounds with their associated biological activities is provided on the DVD.

DVD path: /mglu_x_profiler

REFERENCES

- [1] E. W. Lowe, Jr., M. Butkiewicz, M. Spellings, A. Omlor, and J. Meiler, "Comparative Analysis of Machine Learning Techniques for the Prediction of LogP (Accepted)," presented at the SSCI 2011 CIBCB - 2011 Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Paris, France, 2011.
- [2] J. Edward W. Lowe, Mariusz Butkiewicz, Zollie White III, Matthew Spellings, Albert Omlor, Jens Meiler, "Comparative Analysis of Machine Learning Techniques for the Prediction of the DMPK Parameters Intrinsic Clearance and Plasma Protein Binding," *Proceedings of 4th International Conference on Bioinformatics and Computational Biology 2012*, p. 25, 2012.
- [3] M. Butkiewicz, E. W. Lowe, R. Mueller, J. L. Mendenhall, P. L. Teixeira, C. D. Weaver, and J. Meiler, "Benchmarking Ligand-Based Virtual High-Throughput Screening with the PubChem Database," *Molecules*, vol. 18, pp. 735-756, 2013.
- [4] J. Drews, "Drug discovery: A historical perspective," *Science*, vol. 287, pp. 1960-1964, Mar 17 2000.
- [5] J. H. Van Drie, "Computer-aided drug design: the next 20 years," *Journal of computer-aided molecular design*, vol. 21, pp. 591-601, Oct-Nov 2007.
- [6] R. A. Carr, M. Congreve, C. W. Murray, and D. C. Rees, "Fragment-based lead discovery: leads by design," *Drug Discovery Today*, vol. 10, pp. 987-92, Jul 15 2005.
- [7] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, Jr., "Computational methods in drug discovery," *Pharmacological reviews*, vol. 66, pp. 334-95, Jan 2014.
- [8] C. Hansch, "Use of quantitative structure-activity relationships (QSAR) in drug design (review)," *Pharmaceutical Chemistry Journal*, vol. 14, pp. 678-691, 1980.
- [9] M. A. Lill, "Multi-dimensional QSAR in drug discovery," *Drug Discovery Today*, vol. 12, pp. 1013-1017, Dec 2007.
- [10] J. C. Dearden, "In silico prediction of drug toxicity," *Journal of Computer-Aided Molecular Design*, vol. 17, pp. 119-127, Feb 2003.
- [11] M. Novic and M. Vracko, "Comparison of spectrum-like representation of 3D chemical structure with other representations when used for modelling biological activity," *Chemometrics and Intelligent Laboratory Systems*, vol. 59, pp. 33-44, Nov 28 2001.
- [12] J. Gasteiger, C. Rudolph, and J. Sadowski, "Automatic Generation of 3D-Atomic Coordinates for Organic Molecules," *Tetrahedron Comput. Method.*, vol. 3, pp. 537-547, 1992.
- [13] J. Fang, Y. Dong, G. H. Lushington, Q. Z. Ye, and G. I. Georg, "Support vector machines in HTS data mining: type I MetAPs inhibition study," *J Biomol Screen*, vol. 11, pp. 138-144, 2006.
- [14] D. Hecht and G. Fogel, "High-throughput ligand screening via preclustering and evolved neural networks," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 4, pp. 476-484, 2007.
- [15] G. Harper, J. Bradshaw, J. C. Gittins, D. V. S. Green, and A. R. Leach, "Prediction of biological activity for high-throughput screening using binary kernel discrimination," *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 1295-1300, Sep-Oct 2001.
- [16] T. G. Dietterich, "Machine learning for sequential data: A review," in *Structural, syntactic, and statistical pattern recognition*, ed: Springer, 2002, pp. 15-30.
- [17] R. King, J. Hirst, and M. Sternberg, "New approaches to QSAR: Neural networks and machine learning," *Perspectives in Drug Discovery and Design*, vol. 1, pp. 279-290, 1993.
- [18] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support vector regression machines* vol. 9. Cambridge, MA: MIT Press, 1997.
- [19] J. F. Gunn, 3rd, D. Lester, J. Haines, and C. L. Williams, "Thwarted belongingness and perceived burdensomeness in suicide notes," *Crisis*, vol. 33, pp. 178-81, Jan 1 2012.
- [20] A. J. Smola and B. Schoelkopf, "A tutorial on support vector regression," vol. 14, ed: Kluwer Academic Publishers, 2004, pp. 199-222.
- [21] B. Schoelkopf and A. J. Smola, *Learning with Kernels*. Cambridge, Massachusetts: The MIT Press, 2002.

- [22] D. Winkler, "Neural networks as robust tools in drug lead discovery and development," *Molecular Biotechnology*, vol. 27, pp. 139-167, 2004.
- [23] R. Guha, J. R. Serra, and P. C. Jurs, "Using a Kohonen self-organizing map to generate representative training, cross validation and prediction sets for QSAR modelling.," *Abstracts of Papers of the American Chemical Society*, vol. 226, pp. U448-U448, Sep 2003.
- [24] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, pp. 1464-1480, 1990.
- [25] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, pp. 59-69, 1982.
- [26] D. T. Lee, "On Kappa-Nearest Neighbor Voronoi Diagrams in the Plane," *Ieee Transactions on Computers*, vol. 31, pp. 478-487, 1982.
- [27] R. L. Morin and D. E. Raeside, "A Reappraisal of Distance-Weighted Kappa Nearest Neighbor Classification for Pattern-Recognition with Missing Data," *Ieee Transactions on Systems Man and Cybernetics*, vol. 11, pp. 241-243, 1981.
- [28] J. Meiler and M. Will, "Automated Structure Elucidation of Organic Molecules from ¹³C NMR Spectra using Genetic Algorithms and Neural Networks," *J. Chem. Inf. Comput. Sci.*, vol. 41, pp. 1535-1546, 2001.
- [29] A. Tropsha, "Best Practices for QSAR Model Development, Validation, and Exploitation," *Molecular Informatics*, vol. 29, pp. 476-488, 2010.
- [30] P. Gramatica, "Principles of QSAR models validation: internal and external," *QSAR & Combinatorial Science*, vol. 26, pp. 694-701, May 2007.
- [31] D. M. Hawkins, S. C. Basak, and D. Mills, "Assessing model fit by cross-validation," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 579-586, Mar-Apr 2003.
- [32] G. Piir, S. Sild, and U. Maran, "Comparative analysis of local and consensus quantitative structure-activity relationship approaches for the prediction of bioconcentration factor," *Sar and Qsar in Environmental Research*, vol. 24, pp. 175-199, 2013/03/01 2013.
- [33] J. T. KENT, "Information gain and a general measure of correlation," *Biometrika*, vol. 70, pp. 163-173, April 1, 1983 1983.
- [34] Y.-W. Chen and C.-J. Lin, "Combining SVMs with Various Feature Selection Strategies Feature Extraction." vol. 207, I. Guyon, M. Nikraves, S. Gunn, and L. Zadeh, Eds., ed: Springer Berlin / Heidelberg, 2006, pp. 315-324.
- [35] J. H. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *Ieee Intelligent Systems & Their Applications*, vol. 13, pp. 44-49, Mar-Apr 1998.
- [36] X. Y. Wang, J. Yang, X. L. Teng, W. J. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, pp. 459-471, Mar 1 2007.
- [37] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from Data*, ed: Springer, 1996, pp. 199-206.
- [38] A. Carnero, "High throughput screening in drug discovery," *Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, vol. 8, pp. 482-90, Jul 2006.
- [39] B. A. Posner, "High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics," *Current opinion in drug discovery & development*, vol. 8, pp. 487-94, Jul 2005.
- [40] B. Liu, S. Li, and J. Hu, "Technological advances in high-throughput screening," *American journal of pharmacogenomics : genomics-related research in drug development and clinical practice*, vol. 4, pp. 263-76, 2004.
- [41] M. Butkiewicz, R. Mueller, D. Selic, E. Dawson, and J. Meiler, "Application of Machine Learning Approaches on Quantitative Structure Activity Relationships," Nashville, 2009, pp. 255-262.
- [42] R. Mueller, E. S. Dawson, C. M. Niswender, M. Butkiewicz, C. R. Hopkins, C. D. Weaver, C. W. Lindsley, P. J. Conn, and J. Meiler, "Iterative experimental and virtual high-throughput screening

- identifies metabotropic glutamate receptor subtype 4 positive allosteric modulators," *Journal of molecular modeling*, vol. 18, pp. 4437-46, Sep 2012.
- [43] R. Mueller, A. L. Rodriguez, E. S. Dawson, M. Butkiewicz, T. T. Nguyen, S. Oleszkiewicz, A. Bleckmann, C. D. Weaver, C. W. Lindsley, P. J. Conn, and J. Meiler, "Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening," *ACS chemical neuroscience*, vol. 1, pp. 288-305, Apr 21 2010.
- [44] *PubChem* <http://pubchem.ncbi.nlm.nih.gov>. Available: <http://pubchem.ncbi.nlm.nih.gov/>
- [45] Y. L. Wang, J. W. Xiao, T. O. Suzek, J. Zhang, J. Y. Wang, Z. G. Zhou, L. Y. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte, and S. H. Bryant, "PubChem's BioAssay Database," *Nucleic Acids Research*, vol. 40, pp. D400-D412, Jan 2012.
- [46] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, pp. D1100-D1107, Jan 2012.
- [47] T. I. Oprea, "WOMBAT and WOMBAT-PK: Ten years," *Abstracts of Papers of the American Chemical Society*, vol. 243, Mar 25 2012.
- [48] D. Young, T. Martin, R. Venkatapathy, and P. Harten, "Are the Chemical Structures in Your QSAR Correct?," *QSAR & Combinatorial Science*, vol. 27, pp. 1337-1345, Dec 2008.
- [49] J. Burton, I. Ijjaali, O. Barberan, F. Petitet, D. P. Vercauteren, and A. Michel, "Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset," *J. Med. Chem.*, vol. 49, pp. 6231-6240, 2006.
- [50] D. A. Winkler, "Neural networks as robust tools in drug lead discovery and development," *Molecular biotechnology*, vol. 27, pp. 139-68, Jun 2004.
- [51] D. Plewczynski, M. von Grotthuss, S. A. H. Spieser, L. Rychewski, L. S. Wyrwicz, K. Ginalski, and U. Koch, "Target specific compound identification using a support vector machine," *Combinatorial Chemistry High Throughput Screening*, vol. 10, pp. 189-196, 2007.
- [52] Y. Sakiyama, H. Yuki, T. Moriya, K. Hattori, M. Suzuki, K. Shimada, and T. Honma, "Predicting human liver microsomal stability with machine learning techniques," *Journal of Molecular Graphics and Modeling*, vol. 26, pp. 907-915, 2008.
- [53] F. L. Stahura and J. Bajorath, "Virtual screening methods that complement HTS," *Combinatorial Chemistry & High Throughput Screening*, vol. 7, pp. 259-269, Jun 2004.
- [54] N. Baurin, J. C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot, and L. Morin-Allory, "2D QSAR Consensus prediction for high-throughput virtual screening. an application to COX-2 inhibition modeling and screening of the NCI database," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 276-285, 2004.
- [55] P. J. Conn and J. P. Pin, "Pharmacology and functions of metabotropic glutamate receptors," *Annual review of pharmacology and toxicology*, vol. 37, pp. 205-37, 1997.
- [56] R. Anwyl, "Metabotropic glutamate receptors: electrophysiological properties and role in plasticity," *Brain Research Reviews*, vol. 29, pp. 83-120, 1999.
- [57] M. J. Marino and P. J. Conn, "Modulation of the basal ganglia by metabotropic glutamate receptors: potential for novel therapeutics," *Current drug targets. CNS and neurological disorders*, vol. 1, pp. 239-50, Jun 2002.
- [58] U. C. Campbell, K. Lalwani, L. Hernandez, G. G. Kinney, P. J. Conn, and L. J. Bristow, "The mGluR5 antagonist 2-methyl-6-(phenylethynyl)-pyridine (MPEP) potentiates PCP-induced cognitive deficits in rats," *Psychopharmacology*, vol. 175, pp. 310-8, Sep 2004.
- [59] Y. Zhang, A. L. Rodriguez, and P. J. Conn, "Allosteric potentiators of metabotropic glutamate receptor subtype 5 have differential effects on different signaling pathways in cortical astrocytes," *The Journal of pharmacology and experimental therapeutics*, vol. 315, pp. 1212-9, Dec 2005.
- [60] G. G. Kinney, J. A. O'Brien, W. Lemaire, M. Burno, D. J. Bickel, M. K. Clements, T. B. Chen, D. D. Wisnoski, C. W. Lindsley, P. R. Tiller, S. Smith, M. A. Jacobson, C. Sur, M. E. Duggan, D. J. Pettibone, P. J. Conn, and D. L. Williams, Jr., "A novel selective positive allosteric modulator of

- metabotropic glutamate receptor subtype 5 has in vivo activity and antipsychotic-like effects in rat behavioral models," *The Journal of pharmacology and experimental therapeutics*, vol. 313, pp. 199-206, Apr 2005.
- [61] (2012). *WHO | World Malaria Report 2012*. Available: http://www.who.int/malaria/publications/world_malaria_report_2012/report/en/index.html
- [62] R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay, "The global distribution of clinical episodes of *Plasmodium falciparum* malaria," *Nature*, vol. 434, pp. 214-7, Mar 10 2005.
- [63] A. Dondorp, F. Nosten, K. Stepniewska, N. Day, and N. White, "Artesunate versus quinine for treatment of severe falciparum malaria: a randomised trial," *Lancet*, vol. 366, pp. 717-25, Aug 27-Sep 2 2005.
- [64] A. R. Katritzky, Y. L. Wang, S. Sild, T. Tamm, and M. Karelson, "QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients," *Journal of Chemical Information and Computer Sciences*, vol. 38, pp. 720-725, Jul-Aug 1998.
- [65] M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombar, and A. Tropsha, "Development and Validation of k-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates," *Journal of Medicinal Chemistry*, vol. 46, pp. 3013-3020, 2003.
- [66] H. Geppert, M. Vogt, and J. r. Bajorath, "Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation," *J Chem Inf Model*.
- [67] J. Bajorath, "Integration of virtual and high-throughput screening," *Nat Rev Drug Discov*, vol. 1, pp. 882-894, 2002.
- [68] J. S. Handen, "The industrialization of drug discovery," *Drug Discovery Today*, vol. 7, pp. 83-85, 2002.
- [69] G. Schneider and U. Fechner, "Computer-based de novo design of drug-like molecules," *Nat Rev Drug Discov*, vol. 4, pp. 649-663, 2005.
- [70] R. Perkins, H. Fang, W. Tong, and W. J. Welsh, "Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology," *Environmental Toxicology and Chemistry*, vol. 22, pp. 1666-1679, 2003.
- [71] A. Z. Dudek, T. Arodz, and J. Galvez, "Computational methods in developing quantitative structure-activity relationships (QSAR): a review," *Combinatorial Chemistry & High Throughput Screening*, vol. 9, pp. 213-228, 2006.
- [72] Q. S. Du, R. B. Huang, and K. C. Chou, "Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design," *Current Protein and Peptide Science*, vol. 9, pp. 248-259, 2008.
- [73] C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir, "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients," *Nature*, vol. 194, pp. 178-180, 1962.
- [74] T. Scior, J. L. Medina-Franco, Q. T. Do, K. Martínez-Mayorga, Y. Rojas, and P. Bernard, "How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review," *Current Medicinal Chemistry*, vol. 16, pp. 4297-4313, 2009.
- [75] A. Hillebrecht and G. Klebe, "Use of 3D QSAR models for database screening: a feasibility study," *J. Chem. Inf. Model*, vol. 48, pp. 384-396, 2008.
- [76] R. W. Fawcett, "A radial distribution function analysis of an amorphous calcium phosphate with calcium to phosphorus molar ratio of 1.42," *Calcified Tissue International*, vol. 13, pp. 319-325, 1973.
- [77] M. P. Gonzalez, C. Teran, M. Teijeira, and A. M. Helguera, "Radial distribution function descriptors: an alternative for predicting A2 A adenosine receptors agonists," *European Journal of Medicinal Chemistry*, vol. 41, pp. 56-62, 2006.
- [78] M. P. Gonzalez, Z. Gandara, Y. Fall, and G. Gomez, "Radial Distribution Function descriptors for predicting affinity for vitamin D receptor," *European Journal of Medicinal Chemistry*, vol. 43, pp. 1360-5, 2008.

- [79] B. Hollas, "An Analysis of the Autocorrelation Descriptor for Molecules," *Journal of Mathematical Chemistry*, vol. 33, pp. 91-101, 2003.
- [80] J. Caballero, M. Fernandez, and F. D. Gonzalez-Nilo, "Structural requirements of pyrido[2,3-d]pyrimidin-7-one as CDK4/D inhibitors: 2D autocorrelation, CoMFA and CoMSIA analyses," *Bioorg Med Chem*, vol. 16, pp. 6103-15, 2008.
- [81] J. Caballero, M. Garriga, and M. Fernandez, "2D Autocorrelation modeling of the negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks," *Bioorg Med Chem*, vol. 14, pp. 3330-40, 2006.
- [82] M. Butkiewicz, R. Mueller, D. Selic, E. Dawson, and J. Meiler, "Application of Machine Learning Approaches on Quantitative Structure Activity Relationships (best student paper in IEEE symposium in CIBCB)," presented at the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Nashville, 2009.
- [83] A. Bleckmann and J. Meiler, "Epothilones: Quantitative Structure Activity Relations Studied by Support Vector Machines and Artificial Neural Networks," *QSAR Comb. Sci.*, vol. 22, pp. 719-721, 2003.
- [84] R. Mueller, A. L. Rodriguez, E. S. Dawson, M. Butkiewicz, T. T. Nguyen, S. Oleszkiewicz, A. Bleckmann, C. D. Weaver, C. W. Lindsley, P. J. Conn, and J. Meiler, "Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening," *ACS Chem Neurosci*, vol. 1, pp. 288-305, Apr 21 2010.
- [85] S. Golla, B. J. Neely, E. Whitebay, S. Madihally, R. L. Robinson, Jr., and K. A. Gasem, "Virtual design of chemical penetration enhancers for transdermal drug delivery," *Chemical Biology & Drug Design*, vol. 79, pp. 478-87, Apr 2012.
- [86] H. Sun, S. Shahane, M. Xia, C. P. Austin, and R. Huang, "Structure Based Model for the Prediction of Phospholipidosis Induction Potential of Small Molecules," *J Chem Inf Model*, vol. 52, pp. 1798-1805, 2012/07/23 2012.
- [87] L. Shao, L. Wu, X. Fan, and Y. Cheng, "Consensus Ranking Approach to Understanding the Underlying Mechanism With QSAR," *J Chem Inf Model*, vol. 50, pp. 1941-1948, 2010/11/22 2010.
- [88] K. Simmons, J. Kinney, A. Owens, D. A. Kleier, K. Bloch, D. Argentar, A. Walsh, and G. Vaidyanathan, "Practical Outcomes of Applying Ensemble Machine Learning Classifiers to High-Throughput Screening (HTS) Data Analysis and Screening," *J. Chem. Inf. Model.*, 2008.
- [89] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 1947-1958, 2003.
- [90] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [91] M. Hewitt, M. T. D. Cronin, J. C. Madden, P. H. Rowe, C. Johnson, A. Obi, and S. J. Enoch, "Consensus QSAR Models: Do the Benefits Outweigh the Complexity?," *J Chem Inf Model*, vol. 47, pp. 1460-1468, 2007/07/01 2007.
- [92] R. Hecht-Nielsen, "Counterpropagation networks," *Applied Optics*, vol. 26, pp. 4979-4983, 1987.
- [93] L. M. Patnaik and K. Rajan, "Target detection through image processing and resilient propagation algorithms," *Neurocomputing*, vol. 35, pp. 123-135, 2000.
- [94] A. J. Smola and B. Schoelkopf, *A tutorial on support vector regression* vol. 14: Kluwer Academic Publishers, 2004.
- [95] H. Dug Hun and H. Changha, *Support vector fuzzy regression machines* vol. 138: Elsevier North-Holland, Inc., 2003.
- [96] V. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*: Springer, 1999.
- [97] M. R. H. M. Maruf Hossain, "ROC-tree: A Novel Decision Tree Induction Algorithm Based on Receiver Operating Characteristics to Classify Gene Expression Data," 2006.
- [98] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

- [99] A. P. White and W. Z. Liu, "Technical note: Bias in information-based measures in decision tree induction," *Machine Learning*, vol. 15, pp. 321-329, 1994.
- [100] T. Kohonen, "Self-organization and associative memory," *Self-Organization and Associative Memory, 100 figs. XV, 312 pages.. Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences, volume 8*, vol. 1, 1988.
- [101] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145-1159, 1997.
- [102] R. J. Smith, R. E. See, and G. Aston-Jones, "Orexin/hypocretin signaling at the orexin 1 receptor regulates cue-elicited cocaine-seeking," *Eur J Neurosci*, vol. 30, pp. 493-503, Aug 2009.
- [103] C. J. Winrow, K. Q. Tanis, D. R. Reiss, A. M. Rigby, J. M. Uslaner, V. N. Uebele, S. M. Doran, S. V. Fox, S. L. Garson, A. L. Gotter, D. M. Levine, A. J. Roecker, P. J. Coleman, K. S. Koblan, and J. J. Renger, "Orexin receptor antagonism prevents transcriptional and behavioral plasticity resulting from stimulant exposure," *Neuropharmacology*, vol. 58, pp. 185-94, Jan 2010.
- [104] S. M. Rodems, B. D. Hamman, C. Lin, J. Zhao, S. Shah, D. Heidary, L. Makings, J. H. Stack, and B. A. Pollok, "A FRET-based assay platform for ultra-high density drug screening of protein kinases and phosphatases," *Assay Drug Dev Technol*, vol. 1, pp. 9-19, Nov 2002.
- [105] N. C. Bodick, W. W. Offen, H. E. Shannon, J. Satterwhite, R. Lucas, R. van Lier, and S. M. Paul, "The selective muscarinic agonist xanomeline improves both the cognitive deficits and behavioral symptoms of Alzheimer disease," *Alzheimer Dis Assoc Disord*, vol. 11, pp. S16-22, 1997.
- [106] C. P. Klett and T. I. Bonner, "Identification and characterization of the rat M1 muscarinic receptor promoter," *J Neurochem*, vol. 72, pp. 900-9, 1999.
- [107] A. Medina, N. Bodick, A. L. Goldberger, M. Mac Mahon, and L. A. Lipsitz, "Effects of central muscarinic-1 receptor stimulation on blood pressure regulation," *Hypertension*, vol. 29, pp. 828-34, 1997.
- [108] N. T. Burford and S. R. Nahorski, "Muscarinic m1 receptor-stimulated adenylate cyclase activity in Chinese hamster ovary cells is mediated by Gs alpha and is not a consequence of phosphoinositidase C activation," *Biochemical Journal*, vol. 315, pp. 883-888, 1996.
- [109] B. Arumugam and N. A. McBrien, "Muscarinic antagonist control of myopia: evidence for m4 and m1 receptor-based pathways in the inhibition of experimentally-induced axial myopia in the tree shrew," *Invest Ophthalmol Vis Sci*, vol. 53, pp. 5827-37, Sep 2012.
- [110] W. Wu, R. C. Saunders, M. Mishkin, and J. Turchi, "Differential effects of m1 and m2 receptor antagonists in perirhinal cortex on visual recognition memory in monkeys," *Neurobiol Learn Mem*, vol. 98, pp. 41-6, Jul 2012.
- [111] C. Charlier, N. A. Singh, S. G. Ryan, T. B. Lewis, B. E. Reus, R. J. Leach, and M. Leppert, "A pore mutation in a novel KQT-like potassium channel gene in an idiopathic epilepsy family," *Nat Genet*, vol. 18, pp. 53-5, Jan 1998.
- [112] G. A. Gutman, K. G. Chandy, J. P. Adelman, J. Aiyar, D. A. Bayliss, D. E. Clapham, M. Covarriubias, G. V. Desir, K. Furuichi, B. Ganetzky, M. L. Garcia, S. Grissmer, L. Y. Jan, A. Karschin, D. Kim, S. Kuperschmidt, Y. Kurachi, M. Lazdunski, F. Lesage, H. A. Lester, D. McKinnon, C. G. Nichols, I. O'Kelly, J. Robbins, G. A. Robertson, B. Rudy, M. Sanguinetti, S. Seino, W. Stuehmer, M. M. Tamkun, C. A. Vandenberg, A. Wei, H. Wulff, and R. S. Wymore, "International Union of Pharmacology. XLI. Compendium of voltage-gated ion channels: potassium channels," *Pharmacol Rev*, vol. 55, pp. 583-6, Dec 2003.
- [113] A. S. Dhamoon, S. V. Pandit, F. Sarmast, K. R. Parisian, P. Guha, Y. Li, S. Bagwe, S. M. Taffet, and J. M. Anumonwo, "Unique Kir2.x properties determine regional and species differences in the cardiac inward rectifier K⁺ current," *Circ Res*, vol. 94, pp. 1332-9, May 28 2004.
- [114] G. J. Kaczorowski, O. B. McManus, B. T. Priest, and M. L. Garcia, "Ion channels as drug targets: the next GPCRs," *J Gen Physiol*, vol. 131, pp. 399-405, May 2008.
- [115] H. Sun, X. Liu, Q. Xiong, S. Shikano, and M. Li, "Chronic inhibition of cardiac Kir2.1 and HERG potassium channels by celastrol with dual effects on both ion conductivity and protein trafficking," *J Biol Chem*, vol. 281, pp. 5877-84, Mar 3 2006.

- [116] B. F. Jensen, C. Vind, S. B. Padkjær, P. B. Brockhoff, and H. H. F. Refsgaard, "In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors," *J. Med. Chem*, vol. 50, pp. 501-511, 2007.
- [117] M. T. Nelson, S. M. Todorovic, and E. Perez-Reyes, "The role of T-type calcium channels in epilepsy and pain," *Curr Pharm Des*, vol. 12, pp. 2189-97, 2006.
- [118] M. T. Nelson, P. M. Joksovic, E. Perez-Reyes, and S. M. Todorovic, "The endogenous redox agent L-cysteine induces T-type Ca²⁺ channel-dependent sensitization of a novel subpopulation of rat peripheral nociceptors," *J Neurosci*, vol. 25, pp. 8766-75, Sep 21 2005.
- [119] E. Perez-Reyes, "Molecular physiology of low-voltage-activated t-type calcium channels," *Physiological Reviews*, vol. 83, pp. 117-61, Jan 2003.
- [120] S. M. Ferguson and R. D. Blakely, "The choline transporter resurfaces: new roles for synaptic vesicles?," *Mol Interv*, vol. 4, pp. 22-37, Feb 2004.
- [121] H. Iwamoto, R. D. Blakely, and L. J. De Felice, "Na⁺, Cl⁻, and pH dependence of the human choline transporter (hCHT) in *Xenopus* oocytes: the proton inactivation hypothesis of hCHT in synaptic vesicles," *J Neurosci*, vol. 26, pp. 9851-9, Sep 27 2006.
- [122] Z. Liao, L. Thibaut, A. Jobson, and Y. Pommier, "Inhibition of human tyrosyl-DNA phosphodiesterase by aminoglycoside antibiotics and ribosome inhibitors," *Mol Pharmacol*, vol. 70, pp. 366-72, Jul 2006.
- [123] T. S. Dexheimer, S. Antony, C. Marchand, and Y. Pommier, "Tyrosyl-DNA phosphodiesterase as a target for anticancer therapy," *Anticancer Agents Med Chem*, vol. 8, pp. 381-9, May 2008.
- [124] S. Antony, C. Marchand, A. G. Stephen, L. Thibaut, K. K. Agama, R. J. Fisher, and Y. Pommier, "Novel high-throughput electrochemiluminescent assay for identification of human tyrosyl-DNA phosphodiesterase (Tdp1) inhibitors and characterization of furamide (NSC 305831) as an inhibitor of Tdp1," *Nucleic Acids Res*, vol. 35, pp. 4474-84, 2007.
- [125] M. K. Gilson, H. S. R. Gilson, and M. J. Potter, "Fast Assignment of Accurate Partial Atomic Charges: An Electronegativity Equalization Method that Accounts for Alternate Resonance Forms," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 1982-1997, 2003/11/01 2003.
- [126] J. Gasteiger, C. Rudolph, and J. Sadowski, "Automatic generation of 3D atomic coordinates for organic molecules," *Tetrahedron Computer Methodology*, vol. 3, pp. 537-47, 1990.
- [127] J. Polanski, J. Gasteiger, M. Wagener, and J. Sadowski, "The Comparison of Molecular Surfaces by Neural Networks and its Applications to Quantitative Structure Activity Studies," *Quant. Struct.-Act. Relat.*, vol. 17, pp. 27-36, 1998.
- [128] R. N. Jorissen and M. K. Gilson, "Virtual screening of molecular databases using a support vector machine," *J Chem Inf Model*, vol. 45, pp. 549-61, May-Jun 2005.
- [129] C. W. Yap, C. Z. Cai, Y. Xue, and Y. Z. Chen, "Prediction of torsade-causing potential of drugs by support vector machine approach," *Toxicol Sci*, vol. 79, pp. 170-7, May 2004.
- [130] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev, "Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions," *J Chem Inf Comput Sci*, vol. 43, pp. 2048-56, Nov-Dec 2003.
- [131] V. Sadras and R. Bongiovanni, "Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks," *Field Crops Research*, vol. 90, pp. 303-310, 2004.
- [132] K. Z. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, pp. 629-634, 2004.
- [133] R. Carter and K. N. Mendis, "Evolutionary and Historical Aspects of the Burden of Malaria," *Clinical Microbiology Reviews*, vol. 15, pp. 564-594, 2002.
- [134] K. Marsh, "Malaria disaster in Africa," *The Lancet*, vol. 352, p. 924, 1998.

- [135] N. J. White, "Antimalarial drug resistance," *The Journal of Clinical Investigation*, vol. 113, pp. 1084-1092, 2004.
- [136] C. Wongsrichanalai, A. L. Pickard, W. H. Wernsdorfer, and S. R. Meshnick, "Epidemiology of drug-resistant malaria," *The Lancet Infectious Diseases*, vol. 2, pp. 209-218, 2002.
- [137] D. A. Fidock, P. J. Rosenthal, S. L. Croft, R. Brun, and S. Nwaka, "Antimalarial Drug Discovery: Efficacy Models for Compound Screening," *Nature Reviews Drug Discovery*, vol. 3, pp. 509-520, 2004.
- [138] J. Sachs and P. Malaney, "The economic and social burden of malaria," *Nature*, vol. 415, pp. 680-685, 2002.
- [139] G. F. f. H. Research, "The 10/90 report of research 2001-2002," ed. Global Forum for Health Research, Geneva, 2002.
- [140] S. Nwaka and R. G. Ridley, "Virtual drug discovery and development for neglected diseases through public-private partnerships," *Nature Reviews Drug Discovery*, vol. 2, pp. 919-928, 2003.
- [141] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M.-S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Perlea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. A. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell, "Genome sequence of the human malaria parasite *Plasmodium falciparum*," *Nature*, vol. 419, pp. 498-511, 2002.
- [142] J. M. Carlton, J. H. Adams, J. C. Silva, S. L. Bidwell, H. Lorenzi, E. Caler, J. Crabtree, S. V. Angiuoli, E. F. Merino, P. Amedeo, Q. Cheng, R. M. R. Coulson, B. S. Crabb, H. A. del Portillo, K. Essien, T. V. Feldblyum, C. Fernandez-Becerra, P. R. Gilson, A. H. Gueye, X. Guo, S. Kang/a, T. W. A. Kooij, M. Korsinczky, E. V. S. Meyer, V. Nene, I. Paulsen, O. White, S. A. Ralph, Q. Ren, T. J. Sargeant, S. L. Salzberg, C. J. Stoeckert, S. A. Sullivan, M. M. Yamamoto, S. L. Hoffman, J. R. Wortman, M. J. Gardner, M. R. Galinski, J. W. Barnwell, and C. M. Fraser-Liggett, "Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*," *Nature*, vol. 455, pp. 757-763, 2008.
- [143] M. A. Rush, M. L. Baniecki, R. Mazitschek, J. F. Cortese, R. Wiegand, J. Clardy, and D. Wirth, "Colorimetric High-Throughput Screen for Detection of Heme Crystallization Inhibitors," *Antimicrobial Agents and Chemotherapy*, vol. 53, pp. 2564-2568, 2009.
- [144] R. D. Sandlin, M. D. Carter, P. J. Lee, J. M. Auschwitz, S. E. Leed, J. D. Johnson, and D. W. Wright, "Use of the NP-40 detergent-mediated assay in discovery of inhibitors of beta-hematin crystallization," *Antimicrob Agents Chemother*, vol. 55, pp. 3363-3369, 2011.
- [145] J. Baldwin, C. H. Michnoff, N. A. Malmquist, J. White, M. G. Roth, P. K. Rathod, and M. A. Phillips, "High-throughput Screening for Potent and Selective Inhibitors of *Plasmodium falciparum* Dihydroorotate Dehydrogenase," *The Journal of Biological Chemistry*, pp. 21847-21853, 2005.
- [146] K. A. Kolakovich, I. Y. Gluzman, K. L. Duffin, and D. E. Goldberg, "Generation of hemoglobin peptides in the acidic digestive vacuole of *Plasmodium falciparum* implicates peptide transport in amino acid production," *Molecular and Biochemical Parasitology*, vol. 87, pp. 123-135, August 1997 1997.
- [147] F. J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J. L. Lavandera, D. E. Vanderwall, D. V. S. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon, and J. F. Garcia-Bustos, "Thousands of chemical starting points for antimalarial lead identification," *Nature*, vol. 465, pp. 305-310, 2010.
- [148] D. Plouffe, A. Brinker, C. McNamara, K. Henson, N. Kato, K. Kuhen, A. Nagle, F. Adrian, J. T. Matzen, P. Anderson, T.-g. Nam, N. S. Gray, A. Chatterjee, J. Janes, S. F. Yan, R. Trager, J. S. Caldwell, P. G. Schultz, Y. Zhou, and E. A. Winzeler, "In silico activity profiling reveals the

- mechanism of action of antimalarials discovered in a high-throughput screen," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 9059-9064, 2008.
- [149] R. D. Sandlin, H. M. Carrell, and D. W. Wright, "Hemozoin: Crystal Engineering Survivability," in *Encyclopedia of Inorganic and Bioinorganic Chemistry*, ed: John Wiley & Sons, Ltd, 2012.
- [150] J. M. Combrinck, T. E. Mabothe, K. K. Ncokazi, M. A. Ambele, D. Taylor, P. J. Smith, H. C. Hoppe, and T. J. Egan, "Insights into the Role of Heme in the Mechanism of Action of Antimalarials," *ACS Chemical Biology*, vol. 8, pp. 133-137, 2012.
- [151] K. K. Ncokazi and T. J. Egan, "A colorimetric high-throughput beta-hematin inhibition screening assay for use in the search for antimalarial compounds," *Anal Biochem*, vol. 338, pp. 306-19, Mar 15 2005.
- [152] D. J. Wild and C. J. Blankley, "Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering," *Journal of Chemical Information and Computer Sciences*, vol. 40, pp. 155-162, 2000.
- [153] J. Verma, V. M. Khedkar, and E. C. Coutinho, "3D-QSAR in Drug Design-A Review," *Current Topics in Medicinal Chemistry*, vol. 10, pp. 95-115, 2010.
- [154] J. Meiler, "PROSHIFT: Protein chemical shift prediction using artificial neural networks," *Journal of Biomolecular NMR*, vol. 26, pp. 25-37, 2003.
- [155] I. V. Tetko, V. V. Kovalishyn, and D. J. Livingstone, "Volume learning algorithm artificial neural networks for 3D QSAR studies," *J. Med. Chem*, vol. 44, pp. 2411-2420, 2001.
- [156] Z. Song and N. Roussopoulos, "K-nearest neighbor search for moving query point," *Advances in Spatial and Temporal Databases*, pp. 79-96, 2001.
- [157] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *Systems, Man and Cybernetics, IEEE Transactions on*, pp. 325-327, 1976.
- [158] T. G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," *Machine Learning*, vol. 40, pp. 139-157, 2000.
- [159] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," 1995, pp. 1137-1145.
- [160] D. D. Schoepp, D. E. Jane, and J. A. Monn, "Pharmacological agents acting at subtypes of metabotropic glutamate receptors," *Neuropharmacology*, vol. 38, pp. 1431-76, Oct 1999.
- [161] T. Abe, H. Sugihara, H. Nawa, R. Shigemoto, N. Mizuno, and S. Nakanishi, "Molecular characterization of a novel metabotropic glutamate receptor mGluR5 coupled to inositol phosphate/Ca²⁺ signal transduction," *The Journal of biological chemistry*, vol. 267, pp. 13361-8, Jul 5 1992.
- [162] H. H. Nickols and P. J. Conn, "Development of allosteric modulators of GPCRs for treatment of CNS disorders," *Neurobiology of disease*, vol. 61, pp. 55-71, Jan 2014.
- [163] L. E. Chavez-Noriega, H. Schaffhauser, and U. C. Campbell, "Metabotropic glutamate receptors: potential drug targets for the treatment of schizophrenia," *Current drug targets. CNS and neurological disorders*, vol. 1, pp. 261-81, Jun 2002.
- [164] K. Wisniewski and H. Car, "(S)-3,5-DHPG: a review," *CNS drug reviews*, vol. 8, pp. 101-16, Spring 2002.
- [165] K. J. Gregory, E. N. Dong, J. Meiler, and P. J. Conn, "Allosteric modulation of metabotropic glutamate receptors: structural insights and therapeutic potential," *Neuropharmacology*, vol. 60, pp. 66-81, Jan 2011.
- [166] P. N. Vinson and P. J. Conn, "Metabotropic glutamate receptors as therapeutic targets for schizophrenia," *Neuropharmacology*, vol. 62, pp. 1461-72, Mar 2012.
- [167] J. A. O'Brien, W. Lemaire, M. Wittmann, M. A. Jacobson, S. N. Ha, D. D. Wisnoski, C. W. Lindsley, H. J. Schaffhauser, B. Rowe, C. Sur, M. E. Duggan, D. J. Pettibone, P. J. Conn, and D. L. Williams, Jr., "A novel selective allosteric modulator potentiates the activity of native metabotropic glutamate receptor subtype 5 in rat forebrain," *The Journal of pharmacology and experimental therapeutics*, vol. 309, pp. 568-77, May 2004.

- [168] Y. Chen, C. Goudet, J. P. Pin, and P. J. Conn, "N-{4-Chloro-2-[(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide (CPPHA) acts through a novel site as a positive allosteric modulator of group 1 metabotropic glutamate receptors," *Molecular pharmacology*, vol. 73, pp. 909-18, Mar 2008.
- [169] Y. Chen, Y. Nong, C. Goudet, K. Hemstapat, T. de Paulis, J. P. Pin, and P. J. Conn, "Interaction of novel positive allosteric modulators of metabotropic glutamate receptor 5 with the negative allosteric antagonist site is required for potentiation of receptor responses," *Molecular pharmacology*, vol. 71, pp. 1389-98, May 2007.
- [170] A. S. Hammond, A. L. Rodriguez, S. D. Townsend, C. M. Niswender, K. J. Gregory, C. W. Lindsley, and P. J. Conn, "Discovery of a Novel Chemical Class of mGlu(5) Allosteric Ligands with Distinct Modes of Pharmacology," *ACS chemical neuroscience*, vol. 1, pp. 702-716, Oct 20 2010.
- [171] A. L. Rodriguez, J. C. Tarr, Y. Zhou, R. Williams, K. J. Gregory, T. M. Bridges, J. S. Daniels, C. M. Niswender, P. J. Conn, C. W. Lindsley, and S. R. Stauffer, "Identification of a glycine sulfonamide based non-MPEP site positive allosteric potentiator (PAM) of mGlu5," in *Probe Reports from the NIH Molecular Libraries Program*, ed Bethesda (MD), 2010.
- [172] M. J. Noetzel, J. M. Rook, P. N. Vinson, H. P. Cho, E. Days, Y. Zhou, A. L. Rodriguez, H. Lavreysen, S. R. Stauffer, C. M. Niswender, Z. Xiang, J. S. Daniels, C. K. Jones, C. W. Lindsley, C. D. Weaver, and P. J. Conn, "Functional impact of allosteric agonist activity of selective positive allosteric modulators of metabotropic glutamate receptor subtype 5 in regulating central nervous system function," *Molecular pharmacology*, vol. 81, pp. 120-33, Feb 2012.
- [173] M. Packiarajan, C. G. Ferreira, S. P. Hong, A. D. White, G. Chandrasena, X. Pu, R. M. Brodbeck, and A. J. Robichaud, "Azetidinyloxadiazoles as potent mGluR5 positive allosteric modulators," *Bioorganic & medicinal chemistry letters*, vol. 22, pp. 6469-74, Oct 15 2012.
- [174] M. Packiarajan, C. G. Mazza Ferreira, S. P. Hong, A. D. White, G. Chandrasena, X. Pu, R. M. Brodbeck, and A. J. Robichaud, "N-Aryl pyrrolidinonyloxadiazoles as potent mGluR5 positive allosteric modulators," *Bioorganic & medicinal chemistry letters*, vol. 22, pp. 5658-62, Sep 1 2012.
- [175] J. P. Lamb, D. W. Engers, C. M. Niswender, A. L. Rodriguez, D. F. Venable, P. J. Conn, and C. W. Lindsley, "Discovery of molecular switches within the ADX-47273 mGlu5 PAM scaffold that modulate modes of pharmacology to afford potent mGlu5 NAMs, PAMs and partial antagonists," *Bioorganic & medicinal chemistry letters*, vol. 21, pp. 2711-4, May 1 2011.
- [176] C. M. Niswender, C. K. Jones, and P. J. Conn, "New therapeutic frontiers for metabotropic glutamate receptors," *Current topics in medicinal chemistry*, vol. 5, pp. 847-57, 2005.
- [177] A. L. Rodriguez, Y. Nong, N. K. Sekaran, D. Alagille, G. D. Tamagnan, and P. J. Conn, "A close structural analog of 2-methyl-6-(phenylethynyl)-pyridine acts as a neutral allosteric site ligand on metabotropic glutamate receptor subtype 5 and blocks the effects of multiple allosteric modulators," *Molecular pharmacology*, vol. 68, pp. 1793-802, Dec 2005.
- [178] C. W. Lindsley, D. D. Wisnoski, W. H. Leister, A. O'Brien J, W. Lemaire, D. L. Williams, Jr., M. Burno, C. Sur, G. G. Kinney, D. J. Pettibone, P. R. Tiller, S. Smith, M. E. Duggan, G. D. Hartman, P. J. Conn, and J. R. Huff, "Discovery of positive allosteric modulators for the metabotropic glutamate receptor subtype 5 from a series of N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamides that potentiate receptor function in vivo," *Journal of medicinal chemistry*, vol. 47, pp. 5825-8, Nov 18 2004.
- [179] F. Liu, S. Grauer, C. Kelley, R. Navarra, R. Graf, G. Zhang, P. J. Atkinson, M. Popiolek, C. Wantuch, X. Khawaja, D. Smith, M. Olsen, E. Kouranova, M. Lai, F. Pruthi, C. Pulicchio, M. Day, A. Gilbert, M. H. Pausch, N. J. Brandon, C. E. Beyer, T. A. Comery, S. Logue, S. Rosenzweig-Lipson, and K. L. Marquis, "ADX47273 [S-(4-fluoro-phenyl)-{3-[3-(4-fluoro-phenyl)-[1,2,4]-oxadiazol-5-yl]-piperidin-1-yl}-methanone]: a novel metabotropic glutamate receptor 5-selective positive allosteric modulator with preclinical antipsychotic-like and

- procognitive activities," *The Journal of pharmacology and experimental therapeutics*, vol. 327, pp. 827-39, Dec 2008.
- [180] M. Epping-Jordan, S. Nayak, F. Derouet, H. Dominguez, A. Bessis, E. Le Poul, B. Ludwig, V. Mutel, S. Poli, and R. Addex, "In vivo characterization of mGluR5 positive allosteric modulators as novel treatments for schizophrenia and cognitive dysfunction," in *Neuropharmacology*, 2005, pp. 243-243.
- [181] J. M. Darrah, M. R. Stefani, and B. Moghaddam, "Interaction of N-methyl-D-aspartate and group 5 metabotropic glutamate receptors on behavioral flexibility using a novel operant set-shift paradigm," *Behavioural pharmacology*, vol. 19, pp. 225-34, May 2008.
- [182] M. J. Noetzel, K. J. Gregory, P. N. Vinson, J. T. Manka, S. R. Stauffer, C. W. Lindsley, C. M. Niswender, Z. Xiang, and P. J. Conn, "A novel metabotropic glutamate receptor 5 positive allosteric modulator acts at a unique site and confers stimulus bias to mGlu5 signaling," *Molecular pharmacology*, vol. 83, pp. 835-47, Apr 2013.
- [183] R. Mueller, E. S. Dawson, J. Meiler, A. L. Rodriguez, B. A. Chauder, B. S. Bates, A. S. Felts, J. P. Lamb, U. N. Menon, S. B. Jadhav, A. S. Kane, C. K. Jones, K. J. Gregory, C. M. Niswender, P. J. Conn, C. M. Olsen, D. G. Winder, K. A. Emmitte, and C. W. Lindsley, "Discovery of 2-(2-benzoxazolyl amino)-4-aryl-5-cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 (mGlu(5)): from an artificial neural network virtual screen to an in vivo tool compound," *ChemMedChem*, vol. 7, pp. 406-14, Mar 5 2012.
- [184] T. Noeske, B. C. Sasse, H. Stark, C. G. Parsons, T. Weil, and G. Schneider, "Predicting Compound Selectivity by Self-Organizing Maps: Cross-Activities of Metabotropic Glutamate Receptor Antagonists," *ChemMedChem*, vol. 1, pp. 1066-1068, 2006.
- [185] D. J. Sheffler and P. J. Conn, "Allosteric potentiators of metabotropic glutamate receptor subtype 1a differentially modulate independent signaling pathways in baby hamster kidney cells," *Neuropharmacology*, vol. 55, pp. 419-27, Sep 2008.
- [186] C. M. Niswender and P. J. Conn, "Metabotropic glutamate receptors: physiology, pharmacology, and disease," *Annual review of pharmacology and toxicology*, vol. 50, pp. 295-322, 2010.
- [187] K. J. Gregory, M. J. Noetzel, J. M. Rook, P. N. Vinson, S. R. Stauffer, A. L. Rodriguez, K. A. Emmitte, Y. Zhou, A. C. Chun, A. S. Felts, B. A. Chauder, C. W. Lindsley, C. M. Niswender, and P. J. Conn, "Investigating metabotropic glutamate receptor 5 allosteric modulator cooperativity, affinity, and agonism: enriching structure-function studies and structure-activity relationships," *Molecular pharmacology*, vol. 82, pp. 860-75, Nov 2012.
- [188] eMolecules. (2013, May). Available: www.emolecules.com
- [189] W. P. Walters, M. T. Stahl, and M. A. Murcko, "Virtual screening--an overview," *Drug Discovery Today*, vol. 3, pp. 160-178, 1998.
- [190] J. B. Baell and G. A. Holloway, "New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays," *Journal of medicinal chemistry*, vol. 53, pp. 2719-2740, 2010.
- [191] C. Hansch, D. Hoekman, A. Leo, D. Weininger, and C. D. Selassie, "Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology," *Chem Rev*, vol. 102, pp. 783-812, Mar 2002.
- [192] E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, and W. H. Moos, "Measuring diversity: experimental design of combinatorial libraries for drug discovery," *Journal of medicinal chemistry*, vol. 38, pp. 1431-1436, 1995.
- [193] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106-1114.
- [194] J. Kai, K. Nakamura, T. Masuda, I. Ueda, and H. Fujiwara, "Thermodynamic aspects of hydrophobicity and the blood-brain barrier permeability studied with a gel filtration chromatography," *J Med Chem*, vol. 39, pp. 2621-4, Jun 21 1996.

- [195] M. H. Abraham, W. E. Acree, Jr., A. J. Leo, D. Hoekman, and J. E. Cavanaugh, "Water-solvent partition coefficients and Delta Log P values as predictors for blood-brain distribution; application of the Akaike information criterion," *J Pharm Sci*, vol. 99, pp. 2492-501, May.
- [196] S. Miyamoto and P. A. Kollman, "What determines the strength of noncovalent association of ligands to proteins in aqueous solution?," *Proc Natl Acad Sci U S A*, vol. 90, pp. 8402-6, Sep 15 1993.
- [197] A. Leo, C. Hansch, and D. Elkins, "Partition coefficients and their uses," *Chemical Reviews*, vol. 71, pp. 525-616, 1971.
- [198] R. Wang, Y. Fu, and L. Lai, "A New Atom-Additive Method for Calculating Partition Coefficients," *Journal of Chemical Information and Computer Sciences*, vol. 37, pp. 615-621, 1997.
- [199] R. Mueller, A. L. Rodriguez, E. S. Dawson, M. Butkiewicz, T. T. Nguyen, S. Oleszkiewicz, A. Bleckmann, C. D. Weaver, C. W. Lindsley, P. J. Conn, and J. Meiler, "Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening," *ACS Chemical Neuroscience*.
- [200] M. Butkiewicz, R. Mueller, D. Selic, E. Dawson, and J. Meiler, "Application of Machine Learning Approaches on Quantitative Structure Activity Relationships," presented at the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Nashville, 2009.
- [201] B. Schoelkopf, "SVM and Kernel Methods," *www*, 2001.
- [202] T. Fox and J. M. Kriegl, "Machine learning techniques for in silico modeling of drug metabolism," *Curr Top Med Chem*, vol. 6, pp. 1579-91, 2006.
- [203] J. Meiler, "PROSHIFT: Protein Chemical Shift Prediction Using Artificial Neural Networks," *J. Biomol. NMR*, vol. 26, pp. 25-37, 2003.
- [204] W. P. Walters and M. A. Murcko, "Prediction of 'drug-likeness'," *Adv Drug Deliv Rev*, vol. 54, pp. 255-71, Mar 31 2002.
- [205] I. V. Tetko, V. V. Kovalishyn, and D. J. Livingstone, "Volume Learning Algorithm Artificial Neural Networks for 3D QSAR Studies," *Journal of Medicinal Chemistry*, vol. 44, pp. 2411-2420, 2001.
- [206] J. S. Alex and S. Bernhard, *A tutorial on support vector regression*: Kluwer Academic Publishers, 2004.
- [207] B. C. Drucker H and V. V, "Support vector regression machines."
- [208] C. Yan, J. Hu, and Y. Wang, "Discrimination of outer membrane proteins using a K-nearest neighbor method," *Amino Acids*, vol. 35, pp. 65-73, Jun 2008.
- [209] B. F. Jensen, C. Vind, S. B. Padkjaer, P. B. Brockhoff, and H. H. Refsgaard, "In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors," *J Med Chem*, vol. 50, pp. 501-11, Feb 8 2007.
- [210] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, and J. B. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *J Chem Inf Model*, vol. 46, pp. 2412-22, Nov-Dec 2006.
- [211] S. Ajmani, K. Jadhav, and S. A. Kulkarni, "Three-dimensional QSAR using the k-nearest neighbor method and its interpretation," *J Chem Inf Model*, vol. 46, pp. 24-31, Jan-Feb 2006.
- [212] *MDDR*. Available: <http://accelrys.com/products/databases/bioactivity/mddr.html>
- [213] *Reaxys*. Available: <http://www.reaxys.com/info/>
- [214] A. B. Wagner, "SciFinder Scholar 2006: an empirical analysis of research topic query processing," *J Chem Inf Model*, vol. 46, pp. 767-74, Mar-Apr 2006.
- [215] K. Z. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *IEEE Trans Syst Man Cybern B Cybern*, vol. 34, pp. 629-34, Feb 2004.
- [216] I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?," *Nat Rev Drug Discov*, vol. 3, pp. 711-716, 2004.

- [217] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of health economics*, vol. 22, pp. 151-85, Mar 2003.
- [218] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature reviews. Drug discovery*, vol. 9, pp. 203-14, Mar 2010.
- [219] A. B. Richon, "Current status and future direction of the molecular modeling industry," *Drug Discovery Today*, vol. 13, pp. 665-9, Aug 2008.
- [220] J. A. Williams, R. Hyland, B. C. Jones, D. A. Smith, S. Hurst, T. C. Goosen, V. Peterkin, J. R. Koup, and S. E. Ball, "Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUC_i/AUC) ratios," *Drug metabolism and disposition: the biological fate of chemicals*, vol. 32, pp. 1201-8, Nov 2004.
- [221] S. Schmidt, D. Gonzalez, and H. Derendorf, "Significance of protein binding in pharmacokinetics and pharmacodynamics," *Journal of pharmaceutical sciences*, vol. 99, pp. 1107-22, Mar 2010.
- [222] L. M. Berezhkovskiy, "On the influence of protein binding on pharmacological activity of drugs," *Journal of pharmaceutical sciences*, vol. 99, pp. 2153-65, Apr 2010.
- [223] D. Korolev, K. V. Balakin, Y. Nikolsky, E. Kirillov, Y. A. Ivanenkov, N. P. Savchuk, A. A. Ivashchenko, and T. Nikolskaya, "Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach," *J Med Chem*, vol. 46, pp. 3631-43, Aug 14 2003.
- [224] X. Y. He, J. Wegiel, Y. Z. Yang, R. Pullarkat, H. Schulz, and S. Y. Yang, "Type 10 17beta-hydroxysteroid dehydrogenase catalyzing the oxidation of steroid modulators of gamma-aminobutyric acid type A receptors," *Mol Cell Endocrinol*, vol. 229, pp. 111-7, Jan 14 2005.
- [225] T. Kohonen and P. Somervuo, "Self-organizing maps of symbol strings," *Neurocomputing*, vol. 21, pp. 19-30, 1998.
- [226] B. R. Weil, O. J. Maceneaney, B. L. Stauffer, and C. A. Desouza, "Habitual short sleep duration and circulating endothelial progenitor cells," *Journal of cardiovascular disease research*, vol. 2, pp. 110-4, Apr 2011.
- [227] P. Paixao, L. F. Gouveia, and J. A. Morais, "Prediction of the in vitro intrinsic clearance determined in suspensions of human hepatocytes by using artificial neural networks," *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*, vol. 39, pp. 310-21, Mar 18 2010.
- [228] S. A. Combs, S. L. Deluca, S. H. Deluca, G. H. Lemmon, D. P. Nannemann, E. D. Nguyen, J. R. Willis, J. H. Sheehan, and J. Meiler, "Small-molecule ligand docking into comparative models with Rosetta," *Nature protocols*, vol. 8, pp. 1277-98, 2013.
- [229] G. Lemmon and J. Meiler, "Rosetta Ligand docking with flexible XML protocols," *Methods in molecular biology*, vol. 819, pp. 143-55, 2012.
- [230] H. Wu, C. Wang, K. J. Gregory, G. W. Han, H. P. Cho, Y. Xia, C. M. Niswender, V. Katritch, J. Meiler, V. Cherezov, P. J. Conn, and R. C. Stevens, "Structure of a Class C GPCR Metabotropic Glutamate Receptor 1 Bound to an Allosteric Modulator," *Science*, Mar 6 2014.
- [231] G. C. Liu, Z. C. Lin, and Y. Yu, "Multi-output regression on the output manifold," *Pattern Recognition*, vol. 42, pp. 2737-2743, Nov 2009.
- [232] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199-222, Aug 2004.
- [233] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096-1103.
- [234] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 9999, pp. 3371-3408, 2010.
- [235] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *ESANN*, 2011.

- [236] S. H. Bryant, "PubChem," *Abstracts of Papers of the American Chemical Society*, vol. 230, pp. U1008-U1009, Aug 28 2005.
- [237] J. J. Irwin and B. K. Shoichet, "ZINC - A free database of commercially available compounds for virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, pp. 177-182, Jan-Feb 2005.