Universiteit Leiden
The Netherlands

Proteins in harmony: Tuning selectivity in early drug discovery

Burggraaff, L.

Citation

Burggraaff, L. (2020, October 21). *Proteins in harmony: Tuning selectivity in early drug discovery*. Retrieved from https://hdl.handle.net/1887/136965

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/136965

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University Repository

The handle http://hdl.handle.net/1887/136965 holds various files of this Leiden University dissertation.

Author: Burggraaff, L.
Title: Proteins in harmony: Tuning selectivity in early drug discovery
Issue date: 2020-10-21

Proteins in Harmony
Tuning Selectivity in Early Drug Discovery

Lindsey Burggraaff

The research described in this thesis was performed at the Division of Drug Discovery and Safety of the Leiden Academic Centre for Drug Research (LACDR), Leiden University, The Netherlands.

Cover design and thesis layout by Lindsey Burggraaff

This thesis was printed by Ipskamp Printing

© Lindsey Burggraaff 2020
All rights reserved. No part of this thesis may be reproduced in any form or by any means without permission of the author.

Proteins in Harmony
Tuning Selectivity in Early Drug Discovery

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 21 oktober 2020
klokke 15:00 uur

door

Lindsey Burggraaff
geboren te Purmerend, Nederland
in 1990

Promotors:	Prof. dr. Adriaan P. IJzerman Prof. dr. Herman W.T. van Vlijmen
Co-promotor:	Prof. dr. Gerard J.P. van Westen
Promotiecommissie:	Prof. dr. Hubertus Irth - Voorzitter Prof. dr. Joke A. Bouwstra - Secretaris Prof. dr. Didier Rognan (University of Strasbourg) Dr. Hugo Gutiérrez-de-Terán (Uppsala University) Dr. Michael T.M. Emmerich (Leiden University)

About the title and cover page

Just like a school of fish swim in synchrony, the proteins in the human body behave comparably. Proteins do not act alone, but are directly and indirectly influenced by proteins in their environment. Sequentially, these surrounding proteins are also affected by their neighbors resulting in a beautiful synergy unlocking protein communication. This protein crosstalk is disrupted in disease and can be perceived as a fish swimming against the flow, distressing its shoal. In drug discovery we identify the afflicting fish and steer them into the right direction. Using medicines we can change the behavior of proteins; an undertaking that listens carefully as only the deviant proteins should be targeted. Therefore, this process calls for methods that allow us to tune the selectivity of medicines, enabling the proteins to be in harmony once again.

Contents

Chapter one	General Introduction	9
Chapter two	Advances and Challenges in Computational Target Prediction	21
Chapter three	Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 using Proteochemometric Modeling	49
Chapter four	Quantitative Prediction of Selectivity between the A_1 and A_{2A} Adenosine Receptors	73
Chapter five	Successive Statistical and Structure-Based Modeling to Identify Chemically Novel Kinase Inhibitors	97
Chapter six	Annotation of Allosteric Compounds to Enhance Bioactivity Modeling for Class A GPCRs	129
Chapter seven	General Conclusions	149
	Summary	159
	Samenvatting	163
	Samenvatting voor Leken	167
	Glossary	175
	List of Publications	181
	Curriculum Vitae	185
	Acknowledgements	189

Chapter one

General Introduction

About this thesis

This thesis describes the importance of being able to control the selectivity of potential drug candidates. It explains how computational models are employed to predict and rationalize compound-protein binding and therewith, selectivity of compounds. Moreover, it shows that selectivity can purposely be tuned to target either a single protein or an entire panel of proteins. The challenges of selectivity modeling are addressed based on case studies in the sodium-dependent glucose co-transporters, G protein-coupled receptors, and kinases.

Background

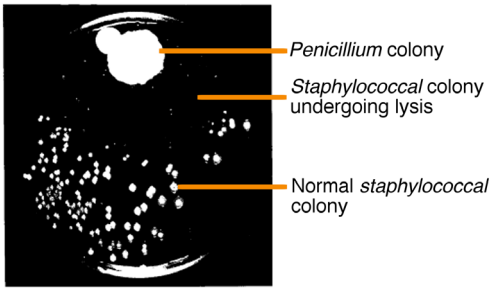
Medicine discovery

The development of new medicines is essential as not every disease can be treated yet and successful treatments that do exist may become ineffective over time due to resistance.[1,2] Moreover, some medications trigger harmful side effects and thus require optimization or development of an alternative therapy.[3] Where in former times medicines were occasionally developed by serendipity,[4] the current era requires a more structured approach to medicine discovery. However, on average it takes about 13 years to develop a new medicine and a total of ~1.8 billion dollar.[5] A typical drug discovery pipeline involves an early drug discovery phase, a preclinical phase, and three subsequent clinical phases. Most of the development costs are made during the (pre-) clinical phases (~1.1 billion dollar) because of failure of drug-candidates due to e.g. unforeseen toxicity or lack of efficacy.[5] By assessing efficacy and toxicity (i.e. side effects) in the early drug discovery phase, issues during the clinical phases may be avoided.

Medicines are chemicals

One can refer to a medicine as many things: a drug, a ligand, a compound, a chemical, and a molecule. Although all of these are correct, the meaning of the words slightly varies and can therefore not always be used interchangeably. To illustrate: all medicines are molecules, but not all molecules are medicines. Although by basis molecules and chemicals do carry the same meaning, they are not always perceived as being similar because of the misconception that naturally-occurring molecules are not chemicals. Furthermore, the terms medicines and drugs

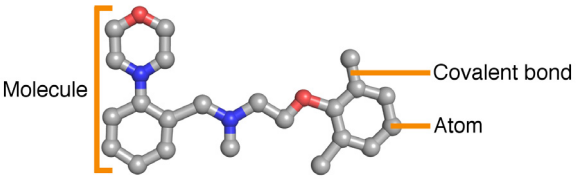
Box 1.1. In 1928, Alexander Fleming discovered the well-known antibiotic penicillin thanks to serendipity. He coincidentally noticed that an unintended mold (penicillium notatum) that was growing on his bacteria cell-culture prevented the bacteria from growing in the space next to the fungus. When he analyzed the fungus more closely he discovered that a substance, which he named penicillin, exhibited anti-bacterial effects that killed the bacteria.[24] *The photograph from A. Fleming (adapted) displays a culture plate showing the dissolution of staphylococcal colonies in the neighbourhood of a penicillium colony.*



can be exchanged naturally as well - although one should be aware that a drug might suggest a stimulant rather than a pharmaceutical. The words ligand and compound are more nuanced and cannot directly be converted into each other. The difference between the two is essential in drug discovery: a compound refers to a chemical that may or may not display any activity, whereas a ligand does exhibit activity. When a ligand has proven its effect *in silico*, *in vitro* or *in vivo*, it may also be referred to as hit compound or lead compound depending on the stage of the process. A ligand that exerts a therapeutic effect and has a good pharmacological profile may eventually become a drug.

The characteristics of compounds are described as physicochemical properties. Such properties include the size, the weight, solubility, and the electrostatic charge of compounds. Furthermore, it describes the number of structural features such as the number of atoms, bonds, rings, and functional groups. The functional groups are important features of compounds as these are able to establish interactions, which include aromatic interactions, interactions through hydrogen bond formation, ionic interactions, or even alteration of covalent bonds.

Box 1.2. A compound is a molecule that is built up by atoms that are connected through covalent bonds. How the atoms are connected through these bonds matters, as different connections result in different molecules. A molecule can contain structural elements such as ring systems: a chain of atoms connected to each other, which results in a ring-shaped construction upon ring closure when the last atom is connected to the first atom of the chain.



Proteins as drug targets

Medicines are molecules. These molecules interact with other, usually bigger, molecules in our bodies, which we call proteins. The human body encompasses over 20,000 different protein-encoding genes, resulting in many proteins that all play a unique role in human physiology.[6] Distortion of a single protein species may already result in pathophysiology, or disease. Early drug discovery focuses on the identification of these pathophysiological proteins and enables the development of compounds that are promising drug candidates. The main protein classes that are targeted by drugs today comprise receptors, enzymes, and transporter proteins. Receptors are responsible for transmission of signals, and encompass the main drug target family: G protein-coupled receptors (GPCRs).[7] It is estimated that about 36% of the current drugs on the market target these receptors, showing their potential in drug discovery.[8] Furthermore, enzymes make up for the second largest drug target class, encompassing 29% of the total amount of drug targets.[7] Some enzymes are responsible for metabolism of compounds, which can be of both natural (endogenous ligands) and synthetic nature (medicines). In contrast to receptors and enzymes, transporters are less explored. Transporters are considered to be more complex, as binding pockets are not always clearly defined. Furthermore, since transporters carry their substrates from inside the cell to the extracellular space or vice versa, they are located within the cell membrane, making it more challenging to elucidate their structure using crystallographic techniques. Consequently, this results in a lack of structural protein information. Although GPCRs are located within the cell membrane as well, many of these protein structures have

been elucidated using a ligand to stabilize the complex. Furthermore, GPCRs have been crystalized by inducing stabilizing mutations and by linking them to easily crystallizable proteins such as T4 lysozyme and apocytochrome b_{562}RIL.[9,10] However, despite of this success for membrane receptors, transporters still remain challenging to crystallize due to instability issues.[11] Nevertheless, upcoming techniques such as cryo-electron microscopy (cryo-EM) may shed light on the structure of transporter proteins as this technique is able to capture proteins whilst avoiding the crystallization issues that occur with commonly used X-ray crystallography.[12]

The three addressed protein classes can be subdivided into protein families. These families are grouped based on protein similarity and function. In general proteins that resemble each other in sequence (**Box 1.3**) and topology (structurally similar), function in a similar matter and are subsequently categorized as family members. Although different proteins may exert a similar function, this does not imply that they are located in the same area. Proteins can be expressed at different locations in the human body, regardless of the location of their family members. Moreover, the same protein may also be expressed in multiple locations. Nevertheless, the location of a protein is usually related to its function: e.g., digestive enzymes are generally present in high amounts in the liver and may occur in lower amounts in the lungs.

Box 1.3. Proteins are constructed from a fixed set of building blocks, which are called amino acids. The human body comprises 20 standard amino acids and one additional non-standard residue named selenocysteine. The residues are linked to each other, creating a chain that facilitates the primary structure of the protein, which is called the protein sequence. The residue type and order of the residues in the sequence define the protein that is formed. When a protein is mutated (e.g., in the case of cancer) one or more residues have been replaced by a different residue type.

Drug-protein interactions

A medicine that acts on a protein makes physical contact with that specific protein: this is called "drug binding". Drugs bind to so-called "binding pockets" on proteins, of which the orthosteric binding pocket is the location where also the natural, or endogenous, ligand or substrate can bind. Any other sites on the protein that allow for ligand binding are called allosteric pockets. When a drug binds to a protein binding pocket, interactions are constituted between the drug and the protein (**Figure 1.1**). These interactions can be of electrostatic, aromatic, or hydrophobic nature, or enable hydrogen bond formation. Moreover, water molecules can additionally form hydrogen bonds with the ligand and the protein. The interactions that are formed between the ligand and the protein are based on the physicochemical properties of both the ligand and the amino acids in the binding pocket. For example, a positive charge on a ligand that is in close proximity to a negatively charged amino acid will result in an electrostatic interaction. In general, an increase in the number of drug-protein interactions improves drug binding. The measure of how well a drug binds to a protein is described as affinity, where a high affinity means good binding and low affinity means poor binding. When a compound is unable to bind to a protein, it has no affinity for that particular protein and is thus not considered a ligand for the respective target.

Binding of a ligand to a protein may result in different effects. Stimulation of proteins can be measured as activation, and blocking as inhibition. In the case of GPCRs, activators are also called agonists and inhibitors antagonists. Additionally, an inverse agonist can deactivate a GPCR and a partial agonist is able to partially activate the receptor instead of triggering full activation as (full) agonists do. The resulting effect depends on the properties of the particular ligand and the protein it is binding to. It is thus possible that a ligand may activate protein A, but inhibits protein

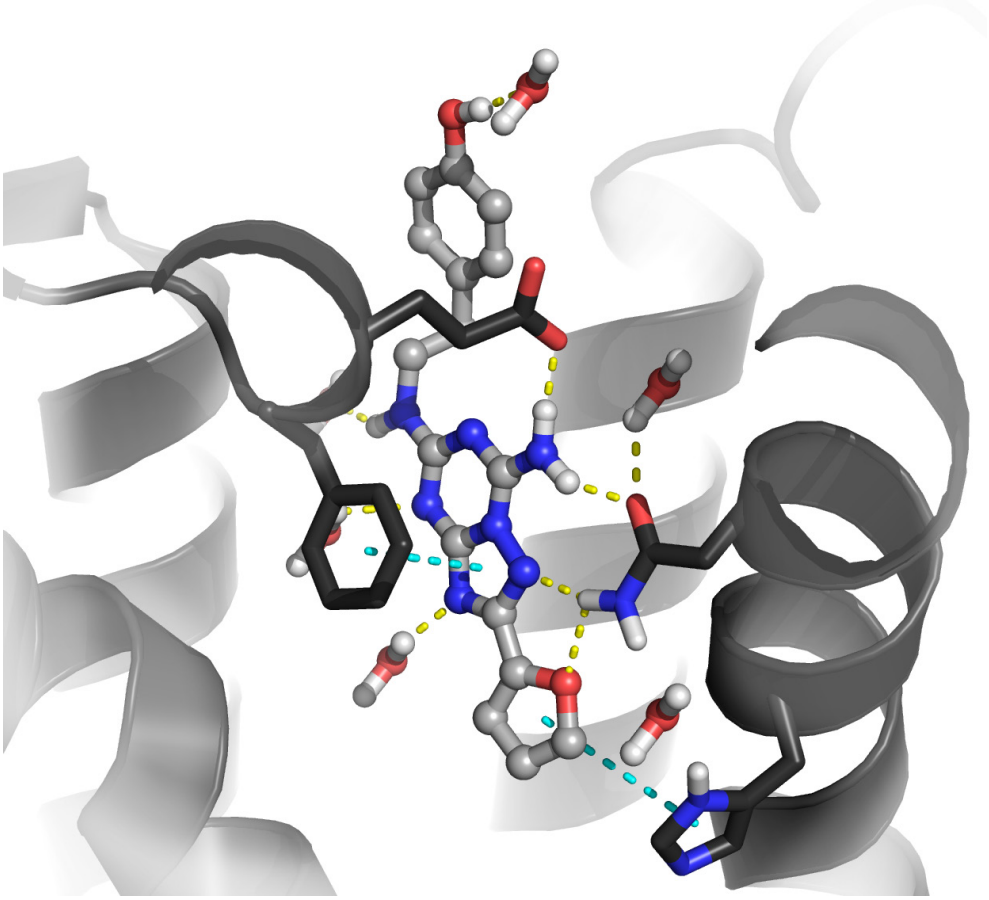


Figure 1.1. A compound (ZM241385) in the binding pocket of the A_{2A} adenosine receptor.[13] Hydrogen bond interactions between the compound (grey), water molecules (red/white), and the protein (black) are displayed in yellow. The aromatic interactions are shown in cyan.

B. Although it might seem contradictory in the case of inhibition, both effects are considered as a result of a compound being bioactive, as in both cases binding of the compound results in a biological effect. The higher the bioactivity of a compound, the stronger is its stimulating or inhibiting effect.

Drugs may compete with natural entities

The properties of a ligand and its target protein are not the only factor influencing drug binding, and therefore its efficacy. When a drug is administered it enters an environment containing many natural ligands and substrates. Some of these natural ligands and substrates may bind to the intended drug target. This may subsequently lead to competition between the drug and the endogenous entity when they both try to bind to the same binding pocket. As a consequence, higher concentrations of the drug, and thus a higher dose, are required to reach the same effect compared to when it would not have to compete. A resolution could be to target the allosteric binding site instead of the orthosteric binding pocket. However, this may face other challenges,

such as tweaking the ability to cross the cell membrane (in case of transmembrane proteins), which may not be feasible for every drug. Moreover, in general orthosteric binding pockets are better explored than allosteric pockets, providing us with less information about allosteric pockets. This lack of data can obstruct the development of certain ligands, as more knowledge may be required to design highly effective drugs.

Selectivity, promiscuity, and polypharmacology

A drug should target the protein that is involved in disease, as targeting of any other protein may result in adverse effects. The more specificly a protein is targeted, the higher the selectivity of a drug. When a drug is not very selective and therefore binds to miscellaneous proteins, it is called promiscuous. Selectivity of a compound can be tuned by changing its molecular structure and properties. For instance, a molecule can be grown into a certain subpocket by adding extra atoms to the molecule. When this subpocket is not present in the anti-target (the protein that should not be bound) the compound does not fit anymore, resulting in achieved selectivity for the on-target over the anti-target. Selectivity of compounds is not limited to specific targeting of only a single protein. In more complex diseases such as cancer, it can be advantageous to target multiple proteins with a single drug to limit the onset of resistance. Multi-target drug design can also be coined as polypharmacology. Although polypharmacological drugs tend to be more promiscuous, selectivity over multiple off-targets can still be of interest. In this case, multiple on-targets can be collectively targeted, while drug binding to off-targets is prevented.

Computers in drug discovery

Computational methods are an important aspect in early drug discovery, as they allow for quick and cost-efficient validation of millions of compounds. Moreover, they can provide insight into ligand-protein binding and guide medicinal chemists in drug design and optimization. Considering that the amount of possible molecules is endless, it would be very inefficient to physically create every molecule and test them all for bioactivity. Even if we would limit the number of molecules to only the medicine-like molecular space by filtering on Lipinski's rule of five[14], still an astonishing 10^{33} compounds would remain.[15] Using computational methods we can create molecules digitally, instead of physically, this saves us a lot of time and resources. Moreover, the compounds can also be screened for bioactivity using the computer in a process called virtual screening. By using computational models, the time to test a potential medicine can be decreased significantly.[16,17] With automated biological assays (high-throughput screening) 10,000 compounds per day can be measured,[18] whereas a relatively slow three-dimensional computational model already is capable of testing ~170,000 compounds per day (24 cores)[i]. For comparison, a traditional (dual-core processor) laptop contains two cores. Faster systems (1500 cores) have reported a virtual screening speed of 115 million compounds per day.[19] A core can be assigned to execute a single task. However, with hyper-threading a core can be split virtually to assign multiple tasks to a single core.

Box 1.4. The Lipinski's rule of five is used to evaluate drug-likeness of compounds. Counter-intuitively the rule includes four guidelines that are based on the physicochemical properties of medicines. Based on these guidelines a compound is considered as a potential drug when: 1) the compound has no more than 5 hydrogen bond donors, 2) the compound has a maximum of 10 hydrogen bond acceptors, 3) the molecular weight of the compound is less than 500 daltons, and 4) the log octanol-water partition coefficient is less than 5.[14]

i Chapter 5 from this thesis

Virtual screening requires models that represent ligand-protein binding. The most comprehensible models contain a three-dimensional representation of the protein (**Figure 1.2**). The affinity of a compound can subsequently be predicted by fitting the compound into the binding pocket of the protein, where a better fit indicates a higher affinity. The three-dimensional protein structures can be derived using crystallography or cryo-EM. With crystallography the protein can be visualized by measuring electron densities of the crystalized protein. The success of the procedure is heavily dependent on the stability of the protein, which can be influenced by the acidity of the solution (pH) that the protein is in, and by co-crystalized factors such as stabilizers and ligands. The stability issue has as a consequence that not all proteins can easily be crystalized. Nevertheless, crystalization of proteins may not be necessary as protein structures may also be derived with cryo-EM, which uses frozen proteins instead. Although this is a viable solution to derive protein structures, resolutions are generally better with crystallography. Proteins for which no protein structure is available (yet) may be structurally modeled if the structure of a close homolog is available. However, if this is not the case, other modeling techniques should be used that do not require information on the protein structure.

Methods based on machine learning or artificial intelligence can predict ligand-protein binding without direct knowledge about the protein structure. In the context of drug discovery, these models are called quantitative structure-activity relationship (QSAR) models and proteochemometric (PCM) models. QSARs are constructed based on compound information only and are commonly used for single targets, whereas PCMs include additional protein sequence data and can encompass multiple proteins in one model. The models rely on statistics and trends in data and are therefore also known as statistical models. Statistical models require a training phase in which they extract trends from data. These trends can subsequently be applied in the test and screening phases to label new compounds. In bioactivity modeling of compounds, models are trained on experimentally measured bioactivities. Therefore, this requires prior knowledge

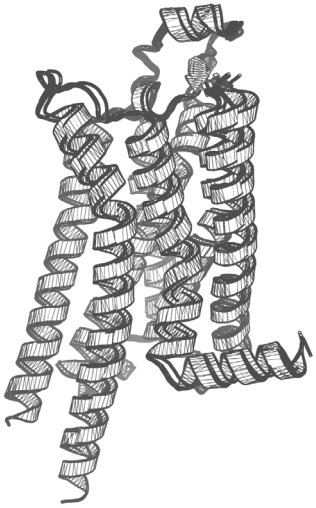


Figure 1.2. Overlay of 3D representations of the A_{2A} adenosine receptor. Protein crystal structures of the exact same protein may vary, as proteins are dynamic and can be influenced by the crystallization procedure. Protein structure PDB codes[20] (pH 5.0-5.5): 4EIY[13], 5IU4[21], 5IU7[21], 5K2C[21], 5NM2[23], 5NM4[23], 5OLG[24], and 5OLZ[24].

on ideally many examples of ligand-protein binding. It is therefore no surprise that lack of data can be one of the major bottlenecks in statistical modeling.

Bioactivity predictions made by computational models are rarely 100% correct. Therefore, it is important to perform experimental validation of the obtained results. This experimental validation commonly includes *in vitro* assessment of the predicted bioactivities of compounds. The hit rate indicates the number of true actives, and thus the rate of good predictions, that results from the *in vitro* biological assay. The expected hit rate can be estimated prior to virtual screening by validation of the computational models with known ligands. These ligands are not used in model training and are therefore considered as new compounds. The better the model is able to retrieve the bioactive compounds, the better the estimated hit rate will be.

Objectives of this thesis

Selectivity of drugs is essential as side effects can be prevented. Therefore, this thesis encompasses multiple approaches to model selectivity in early drug discovery. Additionally, the challenges that are faced in selectivity modeling are discussed. The compound-protein modeling techniques that can be applied in selectivity modeling are thoroughly described in **Chapter 2**. Moreover, this chapter also describes how potential anti-targets can be identified, which allows us to anticipate side effects and enables the design of more selective compounds. **Chapter 3** presents a case study on sodium-dependent glucose co-transporters, where selectivity can be achieved by changing the area of treatment from the kidneys to the small intestine. Additionally, it is shown that novel ligands for these transporters can be found using a machine learning approach in virtual screening. Furthermore, **Chapter 4** describes that machine learning can be applied to model selectivity in the adenosine receptors; a subfamily of the GPCRs. Using this approach, a quantitative measure of selectivity between two proteins can be provided. **Chapter 5** illustrates the concept of polypharmacology in kinases. It presents a workflow in which fifteen kinases are carefully curated and validated, followed by a virtual screen for novel inhibitors using both machine learning and structure-based methods. **Chapter 6** addresses the lack of binding type data (orthosteric or allosteric binder), which influences the performance of computational models that predict compound bioactivities and consequently selectivity. In this chapter text mining is employed to label orthosteric and allosteric compounds. These binding types are subsequently used to improve bioactivity predictions and thus illustrate the necessity of filling data gaps in computational chemistry.

References

(1) Ramos, P.; Bentires-Alj, M. Mechanism-Based Cancer Therapy: Resistance to Therapy, Therapy for Resistance. Oncogene 2015, 34, 3617–3626. https://doi.org/10.1038/onc.2014.314.
(2) Dickey, S. W.; Cheung, G. Y. C.; Otto, M. Different Drugs for Bad Bugs: Antivirulence Strategies in the Age of Antibiotic Resistance. Nat. Rev. Drug Discov. 2017, 16, 457.
(3) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER Database of Drugs and Side Effects. Nucleic Acids Res. 2016, 44, D1075–D1079. https://doi.org/10.1093/nar/gkv1075.
(4) Ban, T. A. The Role of Serendipity in Drug Discovery. Dialogues Clin. Neurosci. 2006, 8, 335–344.
(5) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. Nat. Rev. Drug Discov. 2010, 9, 203–214. https://doi.org/10.1038/nrd3078.
(6) Smith, L. M.; Kelleher, N. L.; Proteomics, T. C. for T. D.; Linial, M.; Goodlett, D.; Langridge-Smith, P.; Ah Goo, Y.; Safford, G.; Bonilla, L.; Kruppa, G.; Zubarev, R.; Rontree, J.; Chamot-Rooke, J.; Garavelli, J.; Heck, A.; Loo, J.; Penque, D.; Hornshaw, M.; Hendrickson, C.; Pasa-Tolic, L.; Borchers, C.; Chan, D.; Young, N.; Agar, J.; Masselon, C.; Gross, M.; McLafferty, F.; Tsybin, Y.; Ge, Y.; Sanders, I.; Langridge, J.; Whitelegge, J.; Marshall, A. Proteoform: A Single Term Describing Protein Complexity. Nat. Methods 2013, 10, 186.
(7) Rask-Andersen, M.; Almén, M. S.; Schiöth, H. B. Trends in the Exploitation of Novel Drug Targets. Nat. Rev. Drug Discov. 2011, 10, 579–590. https://doi.org/10.1038/nrd3478.
(8) Chan, H. C. S.; Li, Y.; Dahoun, T.; Vogel, H.; Yuan, S. New Binding Sites, New Opportunities for GPCR Drug Discovery. Trends Biochem. Sci. 2019, 44, 312–330. https://doi.org/10.1016/j.tibs.2018.11.011.
(9) Chun, E.; Thompson, A. A.; Liu, W.; Roth, C. B.; Griffith, M. T.; Katritch, V.; Kunken, J.; Xu, F.; Cherezov, V.; Hanson, M. A.; Stevens, R. C. Fusion Partner Toolchest for the Stabilization and Crystallization of G Protein-Coupled Receptors. Structure 2012, 20, 967–976. https://doi.org/10.1016/j.str.2012.04.010.
(10) Tate, C. G. A Crystal Clear Solution for Determining G-Protein-Coupled Receptor Structures. Trends Biochem. Sci. 2012, 37, 343–352. https://doi.org/https://doi.org/10.1016/j.tibs.2012.06.003.
(11) Carpenter, E. P.; Beis, K.; Cameron, A. D.; Iwata, S. Overcoming the Challenges of Membrane Protein Crystallography. Curr. Opin. Struct. Biol. 2008, 18, 581–586. https://doi.org/10.1016/j.sbi.2008.07.001.
(12) Shoemaker, S. C.; Ando, N. X-Rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. Biochemistry 2018, 57, 277–285. https://doi.org/10.1021/acs.biochem.7b01031.
(13) Liu, W.; Chun, E.; Thompson, A. A.; Chubukov, P.; Xu, F.; Katritch, V.; Han, G. W.; Roth, C. B.; Heitman, L. H.; IJzerman, A. P.; Cherezov, V.; Stevens, R. C. Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. Science, 2012, 337.
(14) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 2001, 46, 3–26.
(15) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. J. Comput. Aided. Mol. Des. 2013, 27, 675–679. https://doi.org/10.1007/s10822-013-9672-4.
(16) Batool, M.; Ahmad, B.; Choi, S. A Structure-Based Drug Discovery Paradigm. Int. J. Mol. Sci. 2019, 20, 2783. https://doi.org/10.3390/ijms20112783.
(17) Chan, H. C. S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. Trends Pharmacol. Sci. 2019, 40, 592–604. https://doi.org/https://doi.org/10.1016/j.tips.2019.06.004.
(18) Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery-Toxicological Screening Tests. Int. J. Mol. Sci. 2012, 13, 427–452. https://doi.org/10.3390/ijms13010427.
(19) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. Nature 2019, 566, 224–229. https://doi.org/10.1038/s41586-019-0917-9.
(20) Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Res. 2000, 235–242. https://doi.org/10.1093/nar/28.1.235.
(21) Segala, E.; Guo, D.; Cheng, R. K. Y.; Bortolato, A.; Deflorian, F.; Doré, A. S.; Errey, J. C.; Heitman, L. H.; IJzerman, A. P.; Marshall, F. H.; Cooke, R. M. Controlling the Dissociation of Ligands from the Adenosine A2A Receptor through Modulation of Salt Bridge Strength. J. Med. Chem. 2016, 59, 6470–6479. https://doi.org/10.1021/acs.jmedchem.6b00653.
(22) Batyuk, A.; Galli, L.; Ishchenko, A.; Han, G. W.; Gati, C.; Popov, P. A.; Lee, M.-Y.; Stauch, B.; White, T. A.; Barty, A.; Aquila, A.; Hunter, M. S.; Liang, M.; Boutet, S.; Pu, M.; Liu, Z.; Nelson, G.; James, D.; Li, C.; Zhao, Y.; Spence, J. C. H.; Liu, W.; Fromme, P.; Katritch, V.; Weierstall, U.; Stevens, R. C.; Cherezov, V. Native Phasing of X-Ray Free-Electron Laser Data for a G Protein–Coupled Receptor. Sci. Adv. 2016, 2, e1600292. https://doi.org/10.1126/sciadv.1600292.
(23) Weinert, T.; Olieric, N.; Cheng, R.; Brünle, S.; James, D.; Ozerov, D.; Gashi, D.; Vera, L.; Marsh, M.; Jaeger, K.; Dworkowski, F.; Panepucci, E.; Basu, S.; Skopintsev, P.; Doré, A. S.; Geng, T.; Cooke, R. M.; Liang, M.; Prota, A. E.;

Panneels, V.; Nogly, P.; Ermler, U.; Schertler, G.; Hennig, M.; Steinmetz, M. O.; Wang, M.; Standfuss, J. Serial Millisecond Crystallography for Routine Room-Temperature Structure Determination at Synchrotrons. Nat. Commun. 2017, 8, 542. https://doi.org/10.1038/s41467-017-00630-4.

(24) Rucktooa, P.; Cheng, R. K. Y.; Segala, E.; Geng, T.; Errey, J. C.; Brown, G. A.; Cooke, R. M.; Marshall, F. H.; Doré, A. S. Towards High Throughput GPCR Crystallography: In Meso Soaking of Adenosine A2A Receptor Crystals. Sci. Rep. 2018, 8, 41. https://doi.org/10.1038/s41598-017-18570-w.

(25) Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. Influenzæ. Br. J. Exp. Pathol. 1929, 10, 226–236.

Chapter two

Advances and Challenges in Computational Target Prediction

Dominique Sydow[§], Lindsey Burggraaff[§], Angelika Szengel, Herman W. T. van Vlijmen, Adriaan P. IJzerman, Gerard J. P. van Westen, and Andrea Volkamer. Journal of Chemical Information and Modeling. 2019. doi: 10.1021/acs.jcim.8b00832.

[§]these authors contributed equally

Abstract

Target deconvolution is a vital initial step in preclinical drug development to determine research focus and strategy. In this respect, computational target prediction is used to identify the most probable targets of an orphan ligand or the most similar targets to a protein under investigation. Applications range from the fundamental analysis of the mode-of-action over polypharmacology or adverse effect predictions to drug repositioning. Here, we provide a review on published ligand- and target-based as well as hybrid approaches for computational target prediction, together with current limitations and future directions.

Introduction

Target prediction is a key aspect in early preclinical drug development, pivotal to determine the clinical application and to initiate drug development campaigns. For instance, orphan compounds may be known from phenotypic screening, showing changes in cell or organism phenotypes upon compound exposure, without the underlying molecular mechanism being known.[1] Targets for orphan compounds can be experimentally identified with techniques based on chemical proteomics such as affinity chromatography and activity-based protein profiling (ABPP), enabling compound testing against the proteome of cell lysates or even intact cells and organisms.[2–4]

Since these experiments are time and cost extensive, computational alternatives to rapidly predict the primary targets have gained momentum and are commonly known as in silico target prediction, target identification, or target fishing.[5] Herein, a general distinction can be made between ligand-based methods, centered around small molecules, and structure-based methods, implementing information from protein structures.[6] Pivotal to most of these approaches is the chemical similarity principle stating that "similar molecules have a similar biological effect" and conversely that "similar proteins bind similar ligands".[7]

One of the main applications of computational target prediction is to elucidate the mode-of-action of a compound by identifying its potential target. However, the traditional magic bullet paradigm, wherein a ligand has a high potency and selectivity toward a single target, has shifted to the understanding that a ligand affects multiple targets simultaneously.[8,9] In this context, target prediction methods can be used to explore desired polypharmacological effects of ligands to cover disease pathways.[10] Similarly, it can help to spot selectivity or toxicity problems during compound optimization which can potentially lead to unwanted adverse or side effects.[11] Moreover, approved drugs, and hence clinically tested ligands, can be repurposed for different indications if they are also found to interact with a protein target that is part of another disease mechanism.[12–14] This process is called drug repositioning or drug repurposing. Whereas the aforementioned applications focus on predicting targets, computational target prediction methods can also be applied to select ligands that have the highest potential to be relevant chemical probes used for ABPP to characterize the biological function of a poorly understood target.[15–17]

Designed for computational biologists, medicinal chemists, and neighboring disciplines, this review aims to outline the general principle and potential of computational target prediction together with the underlying methods and their application. The article starts with ligand-

based modeling, followed by hybrid approaches (using both ligand and protein data), as well as structure- and interaction-based methods (**Figure 2.1**). Finally, potential pitfalls of the different approaches are covered, and a future perspective is given.

Ligand-based target prediction

Central to ligand-based methods is that they rely on the chemical structure of ligands and associated bioactivity of similar ligands. Ligand-based methods are often used to predict the bioactivity of novel compounds for a specific target (**Figure 2.2**). However, ligand-based methods can also be applied to predict activities for a range of targets. Generally, this can be activity: the target for which the highest activity is predicted is expected to be the most likely target of that query compound.

Typically, the ChEMBL database[18] occasionally in combination with PubChem[19], e.g., in the case of the ExCape database[20], is used as a public source for chemical structures. These databases hold experimentally validated bioactivity data for many compounds tested on a wide range of proteins.

In the following, some general compound descriptors for ligand-based methods are outlined; for specific details, the reader is referred to the review by Rognan.[21] Subsequently, a description of ligand-based methods ordered by increasing complexity coupled to prediction confidence is given (**Table 2.1**). The latter is expected to be higher for the more complex methods.

Compound Descriptors. Compounds in ligand-based models are typically described using their 2D chemical structures. Depending on the data source, an intermediate step can be the conversion from a 1D sequential textual format (e.g., SMILES[22]) to a 2D structure, from which more complex binary vectors such as molecular fingerprints are usually obtained.[23] Different fingerprints are available to describe chemical structures, e.g., atom-pair fingerprints, topological torsion fingerprints, or circular fingerprints, where atom environments are included (e.g., ECFP).[24] Optionally, the 3D shape of compounds is taken into account and translated into similar molecular fingerprints. However, this requires additional information on the 3D conformation of the compounds.[25,26] The use of different chemical fingerprints can impact model performance and was explored by Bender et al.[27] Additionally, physicochemical properties, topological information, and pharmacophore features of compounds can be added as descriptors in a similar way. As a result, each compound is described by an array of numbers forming the compound descriptors. Resemblance between arrays is higher when compounds are more similar to each other.

A more complex representation of compounds, compared to chemical descriptors, are bioactivity spectra descriptors. A spectrum in its simplest form is a binary bitstring representation where each bit represents a protein. Proteins for which a given compound shows activity are marked with a "1" as opposed to those for which this is not the case (marked with "0"). Bioactivity spectra rely on compounds being tested on a range of proteins, instead of compounds being tested on only one or a few targets. Considering compound promiscuity, it is expected that compounds display activity on a number of proteins.[28] Based on the bioactivity spectra, compounds that are not chemically similar but do exert a similar phenotype/bioactivity might be recognized

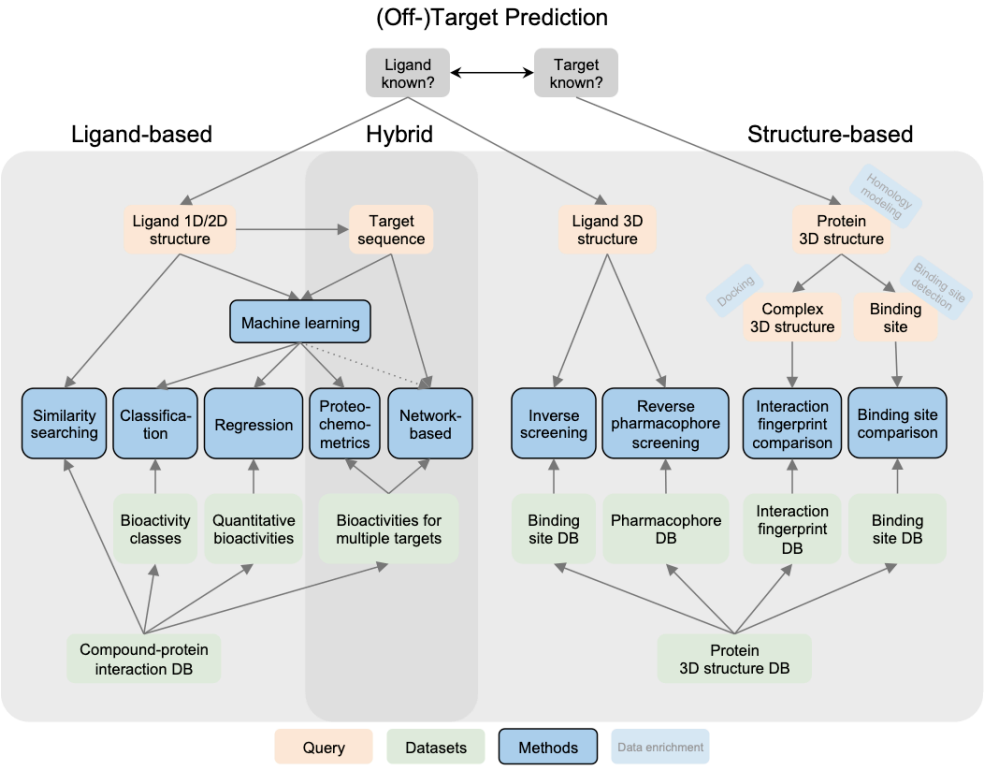


Figure 2.1. Overview of ligand- and structure-based as well as hybrid methods for target prediction (blue) with optional data enrichment strategies (light blue), using database (DB) or training data input (green), separated by applicability depending on available query data (orange). Necessary and potential connections are displayed with solid and dotted arrows, respectively.

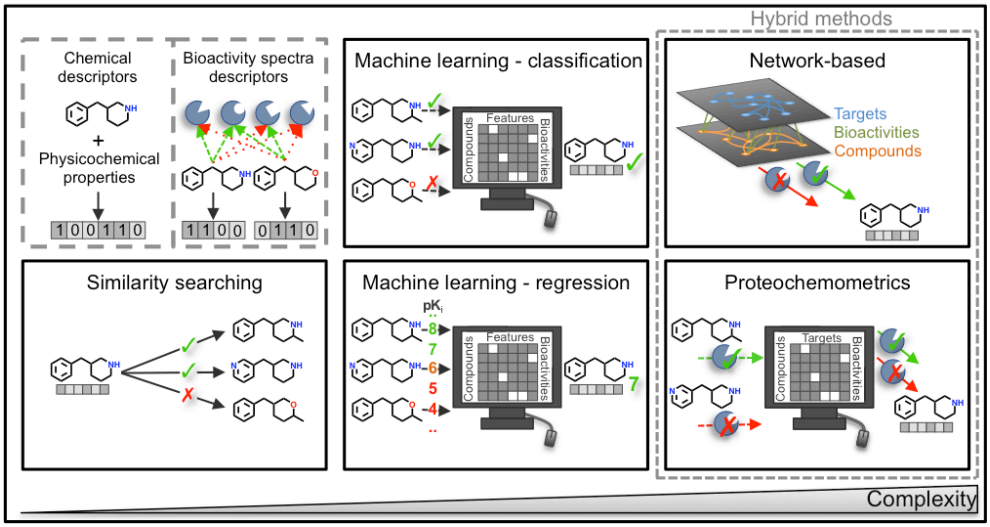


Figure 2.2. Ligand-based methods for target prediction. Descriptors in ligand-based methods are shown in the dashed-lined boxes on the left. Methods increase in complexity from left to right.

Table 2.1. Ligand-Based and Hybrid Methods in Target Prediction*

Name	Data in model training		Training set requirements	Target ranking	Target prediction tools
	Compound	Interaction			
Ligand-based models					
Similarity searching	Chemical structure	-	-	Targets classified based on similarity threshold of compounds	SwissTarget-Prediction,[35] SuperPred,[36] SEA,[40] OCEAN,[45] ROCS,[72] FTrees[73]
Similarity searching	Bioactivities	-	-	Targets classified based on similarity threshold of bioactivity spectra	BASS,[38] BioSEA[4]
Machine learning: Classification	Chemical structure	Activity class	Balanced (in) active classes	Targets classified based on activity class	PIDGIN[74]
Machine learning: Regression	Chemical structure	Bioactivity	Normally/ equally distrib-uted bioactivi-ties	Targets ranked based on bioactivity	-
Hybrid models (ligand- and structure-based)					
Proteoch-emometrics	Chemical structure	Activity class or bioactivity	Balanced (in)ac-tive classes or normally/equal-ly distributed bioactivities	Targets classified or ranked based on bioactivity	ChEMBL mod-els[58,65]
Network-based models	Chemical structure and similarity	Activity class or bioactivity	Sufficient number of con-nections/ bioactivities	Targets classified or ranked based on bioactivity	DINIES,[68] drugCIPHER[69]

*The table gives information on what data is used and how targets are inferred from the model output.

(so-called activity cliffs[29]). Likewise, this bioactivity profile can form an array of numbers that can be implemented as descriptors for similarity searching or machine learning, where activities can be treated as a bioactivity fingerprint. Recently, the biological annotation of compounds has been extended to include gene expression profiles[30,31] and high content cellular images,[32] providing additional, high-dimensional descriptors that can be added to a bioactivity fingerprint in a straightforward way.

Similarity Searching. The simplest and fastest method for target prediction is based on molecular similarity and is often referred to as similarity search or nearest neighbor search.[33] Using a similarity coefficient of choice (e.g., Tanimoto) and any type of compound descriptors (e.g., ECFP), the similarity between a pair of molecules can be quickly generated. For example, finding the most similar 100 compounds for a given query compound in a PubChem-sized library (~96 million compounds) takes a few seconds using chemfp tools developed by Dalke.[34]

The simplest implementation for target prediction based on similarity is to rank the data set compounds based on their similarity toward the query compound and assume that the biologically tested target of the most similar compounds is also the most likely target of the query

compound. Webserver tools that enable the use of this method are, e.g., SwissTargetPrediction[35] and SuperPred.[36] These tools suggest protein targets based on molecular similarity of the query compound to compounds with known bioactivity toward these targets. It should be noted however that these approaches cannot provide a direct quantification of the biological activity of the query compound on the top-ranked targets.

While similarity search is classically performed by comparing chemical descriptors, activity spectra descriptors can also be used (if enough bioactivity data is available). Early work by Kauvar et al.[37] characterized molecular similarity by an affinity fingerprint based on experimental screenings of molecules against a reference panel of selected proteins. Also in BASS[38] (bioactivity profile similarity search), the similarity search is performed based on bioactivity spectra of chemical structures. Here, when the query has experimentally validated activities on a number of targets, additional targets can be predicted based on its bioactivity spectrum. Alternatively, gene expression profiles can be used to predict bioactivities of compounds for targets.[30,39] Both bioactivity spectra and gene expression profiles do not compare the molecular structure of compounds. Therefore, these methods are suited to identify different chemical structures for similar targets.

In contrast to a classical similarity search, similarity ensemble methods are applied to identify targets based on a group of known compounds for that target rather than a single compound. The compounds are first grouped based on interactions (e.g., bioactivity) with the same target(s). The similarity between different compound groups is subsequently calculated, and when defined as being similar, the targets that are known to interact with one compound group are identified as targets for the other compound group(s). The added benefit is that this allows the calculation of statistical measures that can score the relevance of a given retrieved target. When ensemble approaches are applied to identify targets for a query compound, the similarity is measured between this compound and the different compound groups. The targets belonging to the most similar groups are then identified as targets for the query compound. The SEA[40] method utilizes the similarity ensemble concept to group proteins based on ligand topology. Within SEA, the retrieved value is then compared to an expected random value (similar to the way this is implemented in BLAST[41,42]), and subsequently, an "E-value" is returned.[43] This E-value represents the extreme value and indicates the quality of the result. The (similarity) score of the selected samples is compared to what is expected when two random samples are taken into account. E-values closer to zero indicate that it is more unlikely that random samples would have equal similarity as the selected samples. The SEA method has been applied by Lounkine et al.[44] in a target prediction challenge. Here, side effects of 656 compounds were predicted based on compound interactions with 73 off-targets. The results were partially validated by data from hold-out databases or experimentally validated in vitro. Remarkably, off-targets were identified that had very low sequence similarity with the ontarget (e.g., off-target serotonin transporter 5-HTT and ontarget histamine H1 receptor for antihistamine diphenhydramine), indicating that such a ligand-based approach can predict targets without the need of molecular biology information on protein targets. OCEAN[45] is a similar technique, though using different thresholds to determine compound similarities. Finally, BioSEA[46] also applies the same methodology; however, instead of comparing compound similarities based on chemical structure, bioactivity profiles are compared to create ensembles of compounds.

Machine Learning. Similarity search methods consider all features in the compound descriptors as equal. However, statistical methods can weigh the relevance of individual descriptors by connecting them to biological activity of the compounds and are often better suited to extrapolate to new compounds. Machine learning methods require a training phase, which is performed on known active and inactive compounds. Herein, a statistical model is fitted to the data to quantify how chemical descriptors relate to activity. Contrary to the similarity searching example above, this approach returns predicted compound–protein activities rather than a number of compound structures that are similar for a query compound. When applied to a single protein target for a congeneric chemical series, these methods are named quantitative structure–activity relationship (QSAR) models.[47] Given a query compound, QSARs can predict its expected activity based on the compound descriptors. In target prediction, however, more than one protein is considered. Machine learning can both be used for classification (e.g., is the expected affinity higher than a threshold that was defined a priori as active?) or for regression (e.g., what is the predicted K_i value for a compound–protein interaction?). Typically, algorithms such as Random Forest,[48] Support Vector Machines,[49] and Naïve Bayes[50] are applied. However, with more data becoming available and to become more independent of the chosen descriptor, recent work is moving toward deep learning, a method able to directly derive features from molecular structures.[51,52]

An example where machine learning was applied in target prediction is the identification of novel inhibitors for the enzyme mycobacterial dihydrofolate reductase.[53] Here, targets were predicted for a set of query compounds using Naïve Bayesian models. The predicted compound–target interactions were validated in vitro, which indicates the value of such target prediction methods.

Classification. The most frequently used method in ligandbased target prediction is arguably classification.[1,54] Classification requires the setting of an activity threshold for measured interactions to separate the classes. This interaction can be measured binding affinity (e.g., pK_i) but can also be efficacy or other experimental measurements (e.g., pEC_{50}) or even a combination of multiple measurement types (e.g., pChEMBL value).[55] For classification models, a difference can be made between several approaches:

Single Model Multi-Class (SMMC). In this approach, one model is used that predicts the most probable target for a given compound, and target classes are mutually exclusive, in other words a compound cannot be active on more than one target.[56] Given known ligand promiscuity, the SMMC method provides an inaccurate representation of the behavior of ligands and could even be considered to be at odds with the similarity principle.

Ensemble Model Multi-Label (EMML). With EMML, also referred to as ensemble model multi-class, one model is used per protein, and compounds receive a prediction from each model.[1,57] Thus, the sum of protein models where the compound was predicted active on represents the set of potential target proteins. To build the model per protein, all compounds with an activity for the respective protein above a certain threshold are deemed the active class, and all other compounds are typically pooled in the inactive class. For the EMML approach, pooling constitutes a source of error. It might very well be that although a given compound has not been tested on the protein under consideration, it is indeed active yet pooling defines it to be inactive. Thus, potential targets for the query compound may be missed.

Single Model Multi-Label (SMML). Here, one model is used to predict all potential targets for a given molecule, and compounds can belong to multiple target classes (or labels).[56] The active class for a given protein is defined equally as is described for EMML, but all other compounds are not explicitly pooled in an inactive class, merely the ones that were tested to be inactive are considered. A caveat can be that there are none or too few known inactive compounds for good model fitting.

When a query compound is run through a classification model, the output gives the activity class per target (e.g., active/inactive, depending on the previously described approaches and on the predetermined activity threshold). However, regression can directly predict the affinity of a compound.

Pitfalls Defining an "Active" Class. Typically, the activity threshold in classification models is set at 10 μM (i.e., an affinity better than 10 μM defines active interactions, corresponding to a pK_i of 5). This parameter carries a significant influence on effectiveness and applicability of target prediction methods. In principle, for classification, a balanced set of active and inactive compounds is desired. When the activity threshold is set at 10 μM, this gives a skewed distribution of actives and inactives. Recently, target prediction was performed using an affinity value of ~316 nM (corresponding to 6.5 on a logarithmic scale) as the threshold; this leads to a better distribution of active and inactive classes when using ChEMBL data.[58] An added benefit is that this threshold also provides a more relevant prediction of biological activity. Given that the biological error of assays is on average around ~0.5 log units for mixed pKi values, a model using a cutoff of $pK_i = 6.5$ could at worst correspond to an experimental activity of a $pK_i = 6.0$. When a cutoff of $pK_i = 5.0$ (10 μM) is used, this error would be at worst $pK_i = 4.5$ for predicted actives.[57,58] However, the optimal activity threshold for balanced classification sets is dependent on the databases from which compounds and bioactivities are extracted (e.g., ExCape[20] contains more compounds with lower bioactivities than ChEMBL). Furthermore, the targets that are considered can be biased toward reported (in)actives (often in relation to the amount of studies focused on the target, see the Discussion and Future Directions section).

When a reasonable number of inactive compounds is available, but significantly less than the number of active compounds, some workarounds can be applied to train representative models. For instance, active compounds can be divided into smaller subsets in order to train separate models for each subset of actives with the same set of inactives (e.g., random undersampling) and, finally, recombined by ensembling. Ensembling is a technique to combine predictions from multiple models into one prediction that has shown to increase performance.[58,59] The downside of any ensembling method is the unavoidable increase in computational time required as predictions for multiple methods are needed.

Another workaround (which also requires increased computational time) is to construct multiple ligand-based target prediction models at different thresholds (e.g., 10 μM, 1 μM, 100 nM, 10 nM, and 1 nM). However, doing so decreases the available data points for the higher activity thresholds as fewer compounds are known that meet the threshold, and hence, this has a negative effect on the chemical applicability domain. In these cases, regression might allow the use of more data.

Regression. Contrary to classification, regression methods are able to directly train on the strength of a given ligand–protein interaction avoiding the need for a preset threshold. Trained on experimental data, regression models can make quantitative predictions (e.g., K_i values) for compounds based on the chemical structure. These predictions can be directly translated to the interaction (e.g., affinity as a K_i value). Thus, when regression is applied to multiple proteins (using an ensemble of models), the targets can quantitatively be ranked based on predicted compound–protein activity. In addition to predicting activity, the differences in interaction strength for different proteins can be evaluated. Using regression models, the output of a query ligand can constitute a list with ranked targets based on quantitative bioactivity predictions. The output, therefore, does not only define "active" or "inactive" targets but also the activity strength that is reflected by the predicted bioactivity values.

Hybrid methods for target prediction

Similarity searching and machine learning methods-which are classically built on ligand information-can also be applied in more complex systems where protein information is added. Although the underlying mechanism of the methods is the same (e.g., machine learning), the implementation can be different, in turn leading to other application possibilities. This results in alternate methods to model and analyze the data.

Proteochemometrics. With proteochemometrics (PCM), both compound and protein information are combined by addition of an explicit protein descriptor.[60] The most common approach is to add protein information based on knowledge derived from the protein sequence. Sequences are translated into descriptive scores (e.g., z-scales[61]), reflecting the properties of the amino acids of the proteins.[62] Additionally, when structural protein information is available, this may be used to increase descriptor quality as information on binding site location can be included, making the model more accurate compared to using full sequences.[63] PCM can be applied to expand single target models to multiple targets: based on sequence similarity between proteins, data from one protein can be extrapolated to a related one.[64] Another application is increasing the amount of available data (compared to single target models) in order to increase model performance.[63] Several PCM models for target prediction based on ChEMBL data have been reported.[58,65] Such models predict the activities of a query compound for each of the incorporated targets. When these models are based on regression, the most likely target for a query compound can be derived based on the highest predicted activity for that target compared to other targets. Additionally, a quantitative activity score is given per target; therefore, it can be assessed if activity of the query compound for the highest ranked target(s) is sufficient. Noteworthy, as the combination of compound and protein descriptors defines each compound protein pair as a unique pair, even binary class PCM models behave as SMML models. A compound tested to be inactive on protein A can be distinguished from the same compound tested on protein B by the algorithm based on the protein descriptor.

Network-Based Methods. Protein–protein or protein–ligand interactions can be described as a large network similar to a social network. Here, nodes can be proteins, compounds, or both, with the edges being interactions, similarities, or phenotypic effects. These connections can also be weighted based on the strength of interaction (e.g., pK_i). Using chemical structures and similarities between connections, targets can be identified for query compounds.[51] This has led to the publication of several works that use network analysis tools to predict protein

pharmacology.[66,67] Additionally, networkbased target prediction tools such as DINIES[68] and drugCIPHER[69] are made available as open source tools to detect ligand–target interactions for query molecules. The concept of network-based models is often based on similarities between chemical structures but can also include similarities between proteins. More simplistic models implement only one similarity (e.g., protein similarity), whereas more complex models can encompass similarities between protein, chemical structures, and interactions, simultaneously. Such a heterologous network was constituted using three different networks by Chen et al.[70] Here, a protein similarity network (based on sequence similarity) was connected to a compound similarity network by using a ligand–protein interaction network.[71] Therefore, in this network, protein and compound similarities can simultaneously be addressed, which is not possible with only similarity searching as described in the section regarding this topic. Targets for a given query compound can be inferred from the network based on activities (or connections) of similar ligands and their corresponding targets.

Structure-based target prediction

Methods for structure-based target prediction identify the most likely targets for a query ligand or the most similar targets for a query target, using 3D structural, i.e., steric and physicochemical, information (**Figure 2.3**). The former group of approaches focuses on docking a query ligand either to a set of targets (inverse screening) or to a set of pharmacophores inferred from ligand–target complexes (reverse pharmacophore screening), see **Table 2.2**. The latter group of methods compares a query target, either to a set of targets (binding site comparison) or to a set of interactions inferred from ligand–target complexes (interaction fingerprint comparison),[5] see **Table 2.3**.

Typically, the Protein Data Bank (PDB)[75] is used as a public source for protein structures, currently holding more than 140,000 protein structures (accessed in November 2018). Since the binding site is the key to protein function, most methods are proceeded by a binding site annotation step: with a ligand present, binding sites are extracted by a defined ligand-target residue distance cutoff, and without a cocrystallized ligand, binding site detection methods can be invoked.[76] A widely used resource for such annotated binding sites is the scPDB[77] database, containing more than 16,000 ligand-bound binding sites from the PDB and covering about 4700 proteins with 6300 ligands.

Methods for structure-based target prediction are all composed of three main steps, which are described in detail in the individual method paragraphs: (i) binding site encoding, (ii) target screening or comparison, and (iii) target ranking. First, binding sites or ligand-target interactions are encoded using different descriptor techniques and stored in a target database. Second, depending on the method, either a query ligand is screened against the target database, using different docking engines, or a query binding site is compared with the target database, using different similarity measures. Finally, targets are ranked based on a suitable scoring approach.

Inverse Screening. Classically, molecular docking is used to predict both the binding mode and the approximate binding free energy of a set of ligands against one target of interest. In inverse docking, also known as inverse screening or panel docking, this strategy is reversed, and one query ligand is docked to a set of target proteins in order to predict its most likely targets. Most docking tools are theoretically applicable for inverse screening, yet need adaption with respect to

intertarget instead of conventional inter-ligand ranking (**Table 2.2**).[78,79]

(i) Binding Site Encoding. Since the query compound is screened against each target in the data set, the targets need to be preprocessed accordingly. Target databases for methods using conventional docking engines simply contain structure files for binding sites (e.g., TarFisDock[80] and idTarget[81]) or for whole proteins (INVDOCK[82]), preprocessed as required for the respective docking tool. In contrast, iRAISE[83] prepares for an efficient comparison by encoding binding sites with triangle descriptors, which contain pharmacophoric and shape information and are stored as bitmap database, a specialized index for high-dimensional features.

(ii) Target screening. Most inverse screening methods use conventional docking engines, such as DOCK (TarFisDock), MEDock (idTarget), Glide (VTS[84]), or AutoDock Vina (VinaMPI[85] and

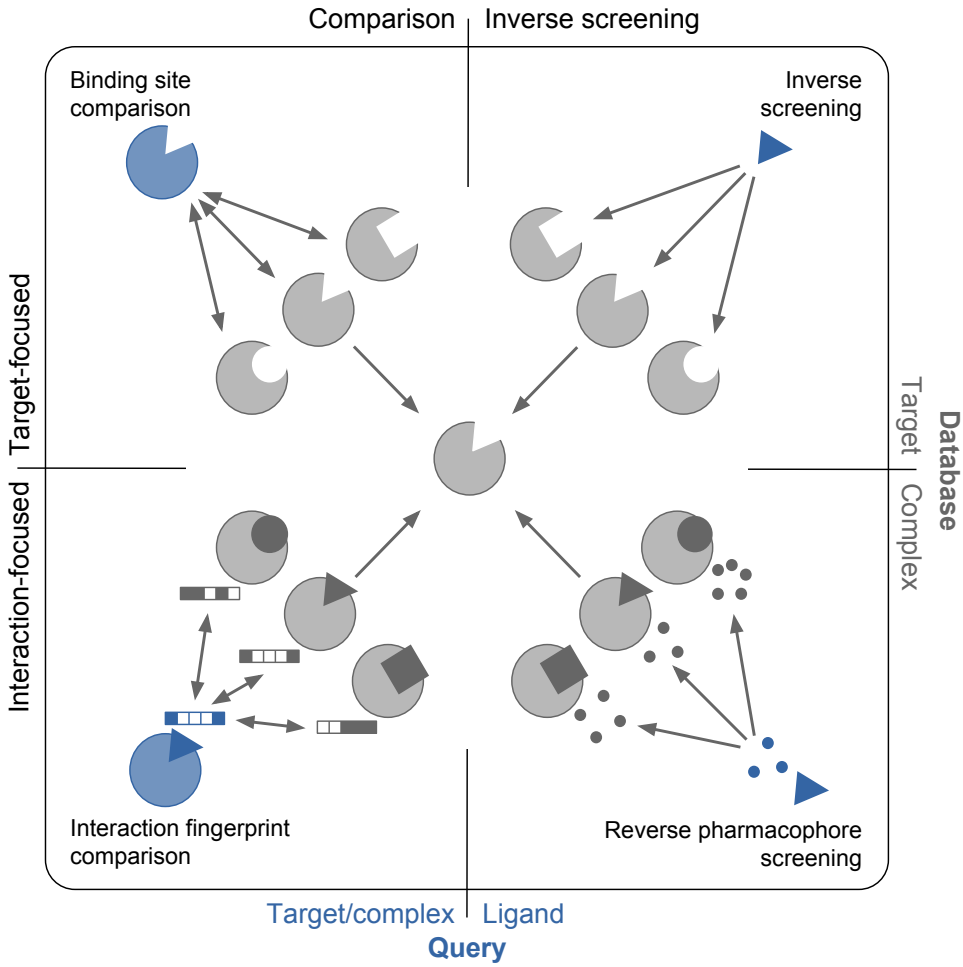


Figure 2.3. Structure-based target prediction: conceptual representation of the four main approaches, i.e. binding site comparison, inverse screening, reverse pharmacophore screening, and interaction fingerprint comparison.

Table 2.2. Structure-based target prediction: Selected methods for inverse screening and reverse pharmacophore screening

Name	Encoding	Target screening		Target ranking	Av.*
		Docking engine	Scoring function		
Inverse screening					
INVDOCK[82]	Sphere-coated surface	DOCK derivative	Interaction energy	-	2
TarFisDock[80]	Sphere-coated surface	DOCK 4.0	Interaction energy	-	2
idTarget[81]	Energetic grid map	MEDock	Binding free energy (AutoDock4 score)	Z-score based on binding free energies of reference complexes	1
VTS[84]	Energetic grid map	Glide	Binding free energy (Glide Gscore)	Gscore comparison to Boltzmannweighted average of reference Gscores	2
VinaMPI[85]	Energetic grid map	AutoDock Vina	Binding free energy (Vina score)	-	1
iRAISE[83]	Bitmap of binned triangles (3 pharmacophore features and cavity shape)	Index-based bitmap comparison	Scoring cascade: clash test, interaction energy and reference cutoff, ligand and pocket coverage	Gaussian-weighted score based on scores for reference complexes	1
Reverse pharmacophore screening					
PharmMapper[90]	Hash table of binned triangles (5 pharmacophore features)	Geometric hashing	Fit score (based on matching feature types and positions)	Z-score based on fit score distribution of reference complexes	1

*Av. = availability: web server, software, or code is (1) free for academic use and/or available upon request or (2) not (yet) available or unclear.

IFPTarget[86]), in order to estimated the fit of the query compound against each protein in the target database. High computational costs are addressed by either parallel screening (VinaMPI and IFPTarget) or by search space reduction. The latter can be realized by aborting the search at the first pose reaching a threshold score based on interaction energies from reference ligand-protein complexes (INVDOCK) or by testing one target representative per precalculated target cluster (based on sequence identity) before screening the entire cluster (idTarget). Usually, energybased functions, such as interaction or binding free energy functions, are used to score the resulting docking poses. In iRAISE, the query ligand is described with triangles, in the same manner as the binding sites before, and is efficiently matched based on bitmap indices, followed by respective superimposition of the ligand and binding site triangles. Finally, iRAISE docking poses are scored using a more extensive approach in the form of a scoring cascade, including a clash test, an interaction energy score, a reference score cutoff (based on the co-crystallized reference ligand), and a ligand and pocket coverage score.

(iii) Target Ranking. Targets are ranked either directly based on the interaction energies of the best docking pose(s) per target (INVDOCK, TarFisDock, and VinaMPI) or based on separate functions tailor-made for inter-target ranking. In the latter approach, each target in the database is profiled beforehand either with a set of ligands using docking (iRAISE and VTS) or with

one co-crystallized ligand (idTarget and IFPTarget). These reference profiles are then used to normalize the scores of docking poses of a query ligand and potential targets.

Inverse screening methods have been widely used for target prediction.[78,79] For example, Scafuri et al.[87] applied idTarget to predict potential targets of apple polyphenols, known for their chemo-preventive effect against colorectal cancer. In a bioinformatics-driven function analysis, the gene expression levels for the predicted targets were shown to be significantly altered in colorectal cancer cells, indirectly linking the investigated apple polyphenols to the predicted targets.

Reverse Pharmacophore Screening. Similar to inverse screening, reverse pharmacophore screening consecutively fits a query ligand in the form of a ligand-based pharmacophore into a precalculated panel of pharmacophore models, derived from protein-ligand complexes. A pharmacophore is defined as an ensemble of physicochemical and steric features that are necessary for the recognition of a ligand by a target, triggering or blocking a biological response.88 Structure-based approaches derive such pharmacophores from a target complex, whereas ligand-based pharmacophores consider solely ligand properties. Several studies have conducted reverse pharmacophore screening for polypharmacology, using available standard software packages that allow for rapid pharmacophore model building and evaluation.[89] However, to the knowledge of the authors, the only available automated workflow for pharmacophore-based target prediction is PharmMapper.[90]

In PharmMapper, the interactions of selected ligand-target complexes are encoded as pharmacophore feature triplets, stored in a hash table, and deposited in a target database (i). For target screening (ii), ligand-based pharmacophores are generated for multiple conformations of the query ligand. Each conformer pharmacophore is described in form of triplets and aligned onto each pharmacophore triplet in the target database, using triangle hashing. Subsequently, targets are scored based on the overlap of feature types and positions between the ligand and target pharmacophores. Finally, each target score is normalized by a reference score for target ranking (iii). The reference score per target reflects the score distribution of matching all ligand pharmacophores extracted from the original protein-ligand complex structures in the database against the target pharmacophore.

Reverse pharmacophore screening was often applied to search for targets of compounds in Chinese traditional medicine (CTM).[79] For example, Liu et al.[91] used PharmMapper to predict the glucocorticoid receptor, p38 mitogen-activated protein kinase, and dihydroorotate dehydrogenase as potential targets of berberine, a compound used in CTM to treat cancers including melanoma. Experimental tests confirmed the predicted targets to be potentially involved in the anti-melanoma effect of berberine.

Binding Site Comparison. Target comparison is based on the assumption that similar proteins - or more precisely binding sites - bind similar ligands. Various binding site comparison methods have been developed, pursuing different strategies to encode binding sites, as well as to measure and score their similarities[92,93] (**Table 2.3**).

(i) Binding Site Encoding. The structural complexity of binding sites is reduced to labeled representatives, whose spatial arrangement is encoded and stored in a database, to be compared

Table 2.3. Structure-based target prediction: Selected methods for binding site and interaction fingerprint comparison[a]

Name	Encoding			Comparison	Scoring	Av
	Representatives	Label	Pattern			
Binding Site Comparison *Alignment-Based Methods*						
SiteBase[104]	[1]Atoms	5 atom types	[5]On-the-fly triangles	Geometric matching	Matching atoms	2
(Med-) SuMo[95,96]	[1]Per-residue: pseudocenters (PCs)	Chemical groups	[5]Triangles as graph of adjacent triangles	Geometric matching, stepwise connection of adjacent matches	Size of connected matches	1,3
SiteEngine[101]	[1]Per-residue: PCs & surface patches	5 pharmacophoric features	[5]Triangles in hash table	Geometric hashing	Matching surface patches & PCs	1
CavBase[94]	[1]Per-residue: PCs & surface patches	5 pharmacophoric features	[4]Graph	Clique detection	Matching surface patches & PCs	3
eF-site[113]	[2]Triangulated surface	Electrostatics & surface curvature	[4]Graph	Clique detection		1
ProBis[97]	[2]Triangulated surface	5 pharmacophoric features	[4]Subgraphs	Clique detection (per subgraph)	Matching surface patches & residues	1
PoLiMorph[114]	[3]Grid-based cavity volume	5 physicochemical features*	[4]Fuzzy graph/self-organizing map	Error-tolerant graph matching	Matching vertices	2
Alignment-Free Methods						
PocketMatch[115]	[1]Per-residue: C_α, C_β & centroid = A	5 amino acids groups = B	[6]90 distance histograms for all A-B combinations	Corresponding histogram comparison	Average matching distance bins	2
RAPMAD[103]	[1]Per-residue: PCs p_i incl. 2 references p_1 & p_2	7 pharmacophoric features: 7 PC subsets s_i	[6]14 distance histograms: p_1-p_i and p_2-p_i per s_i	Corresponding histogram comparison	Jensen-Shannon divergence	2
FuzCav[99]	[1]Per-residue: C_α	6 pharmacophoric features	[5]Triangles as 4833 int fingerprint	Fingerprint comparison	Matching non-zeros	1
KRIPO[116]	[1]Defined points relative to residue	5 pharmacophoric features	[5]Triangles as fuzzy fingerprint(s)	Fingerprint comparison	Modified Tanimoto index	1
PocketFEATURE[117]	[1]Per-residue: center with 6 shells = microenvironment (ME)	80 physicochemical features* per shell	[7]480 int fingerprints per ME	Fingerprint comparison per ME pair if same amino acid	Sum of bestscoring (Tanimoto index) ME pairs	2

Name	Encoding			Comparison	Scoring	Av
	Representatives	Label	Pattern			
Accelerated Alignment-Based Methods						
BSAlign[105]	[1]Per-residue: PCs	5 physicochemical features*	4Reduced (red.) graph	Clique detection, red. product graph	Matching residues & RMDS	1
SiteAlign[106]	[2]80-fold triangulated sphere at binding site center	8 topological and chemical descriptors mapped to triangles	[5]640 int fingerprint	Fingerprint comparison	Average of normalized triangle differences	1
TrixP[102]	[1]Pharmacophoric feature points	3 pharmacophoric features	[5]Triangles in bitmap	Index-based bitmap comparison	Matching triangles	2
FLAP[100]	[2]Clustered GRID-MIFs	5 pharmacophoric features	[5]Quadruplets as 11 int fingerprints	Fingerprint comparison	Matching quadruplets	3
BioGPS[118]	[2]Clustered GRID-MIFs	3 pharmacophoric features	[5]Quadruplets as 11 int fingerprints	Fingerprint comparison	MIF volume overlap per feature	2
Volsite/Shaper[98]	[3]Grid-based cavity volume	7 pharmacophoric features	[7]Volume as smooth Gaussian function	Volume overlap	Matching pharmacophoric features	1
Interaction Fingerprint Comparison						
SIFt[110]	Interacting residues	7 pharmacophoric features	Per-residue: 7 bit vector, concatenated in fixed order	Fingerprint comparison	Tanimoto index	2
TIFP[111]	Pseudoatoms between interacting ligand-target pairs	7 pharmacophoric features	Triplets as 210 int fingerprint	Fingerprint comparison	Tanimoto index	2
SPLIF[112]	Interacting fragments	Atom and bond types	ECFP2 fingerprint	RMSD for all matching fingerprint bits	Matching ligand and protein atoms	2
LIFt[109]	Atom-by-atom ligand-target interactions	10 pharmacophoric features	Interaction fingerprint	Fingerprint comparison	Tanimoto index	2
IFPTarget[86]	Interacting residues	8 pharmacophoric features	Label × residue matrix	Matrix comparison between query and reference complexes	Modified Tanimoto index & energybased score for query complex	2

[a]Binding sites are encoded based on [1]per-residue points, binding site [2]surfaces, and [3]volume, and are represented as [4]graph, [5]triangles/quadruplets (e.g. binned into fingerprints/hash tables/bitmaps), [6]distance distributions of atom pairs, and [7]volume functions. RMSD = root-mean-square deviation; MIFs = molecular interaction fields; Av. = availability: web server, software, or code is (1) free for academic use and/or available upon request, (2) not (yet) available or unclear, or (3) commercially available; *Including pharmacophoric and additional features, e.g. buriedness.

with a query binding site encoded accordingly. Binding site representatives can be per-residue points (e.g., CavBase[94] or (Med-)SuMo[95,96]), binding site surfaces (e.g., ProBis[97]), or binding site volumes (e.g., Volsite/Shaper[98]), with labels mostly containing pharmacophoric information. The spatial arrangement of these representatives is often encoded as graphs (e.g., CavBase) and triangles/quadruplets. The latter are binned by their edge lengths and vertex labels and stored as fingerprints (e.g., FuzCav[99] and FLAP[100]), hash tables (SiteEngine[101]), or bitmaps (TrixP[102]), whereas (Med-) SuMo[95,96] uses a graph of adjacent triangles. Alternate methods describe binding sites as distance distributions between aforementioned per-residue points (e.g., RAPMAD[103]), or with volume functions (Volsite/Shaper).

(ii) Binding Site Similarity Measure. Common strategies for measuring binding site similarities can be divided into alignment-based (often slower) and alignment-free methods (mostly faster), as well as accelerated alignment-based methods. The latter combine the speed of alignment-free methods with the visual interpretability of alignment-based methods. Alignment-based methods calculate and perform the best possible structural superimposition of two binding sites based on their encoded features, using geometric matching and hashing of two triangle sets (e.g., SiteBase[104] and SiteEngine, respectively) or most commonly clique detection between two graphs (e.g., CavBase). The latter approach searches the maximum complete subgraph (clique) in a product graph, which is built from a target and query graph with matching vertices and edges. Many alignment-free methods operate on the comparison of fingerprints (e.g., FuzCav) or of distance histograms (e.g., RAPMAD). Accelerated alignment-based methods use efficient data structures for rapid comparison, with subsequent binding site alignments for scoring and visual interpretation. Those methods include strategies to reduce graph complexity before clique detection (BSAlign[105]), to compare binding site volumes using smooth Gaussian functions (Volsite/Shaper), and to store binned 3-point pharmacophores in bitmap indices (TrixP). Moreover, properties of a binding site can be projected to a triangulated sphere positioned at its center, stored as fingerprint to be iteratively compared, and aligned to another binding site fingerprint (SiteAlign[106]).

(iii) Binding Site Similarity Ranking. Alignment-based methods score the similarity of binding sites based on the mutual overlap and/or root-mean square deviation (RMSD) of their associated encoded features. In contrast, alignment-free methods mainly calculate fingerprint similarity based on the number of matching fingerprints, if multiple fingerprints exist per binding site (e.g., FLAP), or based on the Tanimoto coefficient, if only one fingerprint per binding site (e.g., FuzCav) is calculated.

An exemplary application of binding site comparison is a study on cross-reactivity using SiteAlign by De Franchi et al.[107] Virtual screening of Pim-1 kinase against ATP-binding sites showed high similarity to synapsin I, a protein regulating neurotransmitter release in the synapse, suggesting a crossreaction of protein kinase inhibitors with synapsin I. Biochemical validation revealed nanomolar affinities for pankinase inhibitor staurosporine and selective Pim-1 kinase inhibitor quercetagetin for synapsin I. These findings were proposed as possible explanations for the observed downregulation of neutrotransmitter release by some protein kinase inhibitors.

Interaction Fingerprint Comparison. Interaction fingerprints (IFPs), or protein-ligand fingerprints, are vectors that encode information on interacting ligand and target moieties, such as hydrogen bond, hydrophobic, charge, aromatic, and metal-binding interactions. IFPs are often used in combination with screening methods in order to rescore docking poses.[108] Only a few IFP-based pipelines have been published for target prediction so far. Note that they require a ligand placement step for IFP calculation. Thus, for IFP encoding (i), the query ligand has to be docked against the target structure(s). Generally, IFP methods either map detected interactions to ligand atoms (e.g., LIFt[109]), to target binding site residues (e.g., SIFt[110] and IFPTarget[86]), or define a ligand- and targetindependent fixed length fingerprint (e.g., TIFP[111] and SPLIF[112]). Similar to the alignment-free fingerprint-based binding site comparison, the comparison of two IFPs is usually based on the Tanimoto coefficient (ii), and targets are rankordered accordingly (iii). In the following, two tools are introduced: In the first approach, interactions are mapped on the ligand; thus, ligand IFPs are compared. In the second, information is mapped on the target residues, and subsequently, target IFPs are compared.

Cao and Wang[109] propose a pipeline for off-target prediction exemplified on a tubulin agent with kinase-cross activity. The tubulin agent complex structure is the starting point to generate the ligand-based interaction fingerprint (LIFt) for the query compound. Next, the query ligand is docked to a panel of kinase structures. The best-scoring pose per ligand-kinase complex is encoded as LIFt, documenting interactions per ligand atom. Finally, these predicted panel LIFts are compared (Tanimoto coefficient) to the known reference LIFt and ranked accordingly. In contrast, IFPTarget by Li et al.[86] first sets up a target database, where the co-crystallized ligand is used to define the reference target IFP, documenting per-residue interactions. Next, the query ligand is docked to the same panel of targets, and the top-scoring pose for each target is used to generate the docked target IFP. Subsequently, reference and docked target IFPs are compared and ranked by a final score that integrates aforementioned energy-based docking and IFP-based scores.

The presented methods are strongly intertwined with a docking (inverse screening) procedure: Two IFPs can only be compared if they have one constant component (LIFT: same ligand in two different structures, or IFPTarget: same structure with two different ligands) because otherwise the IFP lengths and order differ. Here, the third category of ligand and protein invariant fingerprints, such as TIFP by Desaphy et al.,[111] could find a remedy, but has, to the knowledge of the authors, not yet been used for target prediction.

Consideration of Target Flexibility in Structure-Based Methods. Proteins are flexible, existing in transient conformational states, whereby only a subset may be receptive to ligand binding. Such flexibility is to some extent implicitly considered by the coarse-grained representation of binding sites in the encoding step, such as binned distances (e.g., RAPMAD and FuzCav) and fuzzified graphs (PoLiMorph[114]), as well as by including tolerances during the matching step. Small side-chain flexibility can be explicitly included by, e.g., representing rotatable hydrophilic interactions (TrixP) or "on-the-fly" conformational sampling of side chains (FLAP and BioGPS[118]). Instead of conformational sampling, different parts of the binding site can be investigated separately from each other in order to spot local similarities. Some methods therefore allow for partial shape matching (TrixP) or local examination of binding site segments (ProBis). Inverse screening methods usually treat the target structure as rigid body, while considering

ligand flexibility by conformational sampling of the ligand (e.g., iRAISE and INVDOCK).

However, information on protein flexibility can be enriched by including protein ensembles in screening databases, either derived from a set of experimentally determined structures or from molecular dynamics (MD) simulations. The former approach is to some extent integrated whenever methods are built upon a database containing multiple structures per protein (e.g., scPDB-based target databases); however, so far, those structures have not been statistically evaluated as one protein ensemble. Furthermore, such PDB-derived protein ensembles can only cover protein classes with high coverage. Methods describing binding site changes based on MD simulations, as described in TRAPP[119] for transient pockets, are already available but have not been integrated yet into a workflow for target prediction.

Discussion and future directions

Since without sufficient data computational target prediction would not be possible at all, we first discuss the beauty and peril of current data sources. We then cover challenges in target ranking and method validation as well as directions on how to overcome them.

Data. Usage of in silico techniques for target prediction has been enabled in the first place by the rapidly increasing amount of available structural, chemical, and biological data. In this respect, the increasing availability of open access databases for drug discovery should be appreciated, with the PDB,[75] ChEMBL,[18] PubChem,[19] and DrugBank[120] databases being arguably the most well known. While the speed of computation has increased at a phenomenal rate with transistor counts roughly doubling every two years[121] (slowing down in recent years[122]), data availability and quality still form the bottleneck.[20,123] Given more data, more intricate methods can be applied, which should result in higher quality predictions.[21] This does not only concern bioactivity data but also structural information on proteins.[75]

In *ligand-based methods*, the large amount of available bioactivity data is used for model training. Lack of data here typically means that there are not enough experimentally derived activities of compounds for a given target. One way to overcome this is using computational target prediction to fill in the expected bioactivities for proteins that were not experimentally tested.[54,124] However, even if sufficient data is available, this does not directly mean the data quality is adequate. It has been shown that the experimental error in bioactivity databases can be substantial.[33,125] In public data, experimental activities are not derived following the same standard operating procedure or are even from the same lab or assay. This leads to a relatively large experimental error in the data (on average 0.47 log units for mixed pK_i data),[33] which is reflected in the prediction accuracy of the models. Data quality and bias each determine the applicability domain of a model and should therefore be addressed early on by comparing the similarity between training and screening compounds. For instance, models trained on smaller or more hydrophobic molecules may not be able to make reliable predictions for larger or more hydrophilic compounds. Furthermore, high chemical similarity within the training set leads to a bias toward a similar group of compounds. Therefore, a wide diversity in chemical space is more favorable than a large compound set encompassing a congeneric series of ligands. Models trained only on close analogues cannot predict activities of very dissimilar compounds reliably. In summary, in order to build reliable models, important factors to check are the amount of data and heterogeneity (as discussed here),

as well as the bias toward (in)actives (see Pitfalls Defining an "Active" Class section) and toward certain targets (see Target Ranking section).

Structure-based methods build on the structural arrangement of binding site atoms, experimentally derived from currently mostly X-ray crystallography. Such structural arrangements are (i) less reliable with decreasing resolution and (ii) represent only a static (and maybe even artificial) conformational state. The former is usually addressed with resolution thresholds (e.g., < 3 Å in case of the scPDB), whereas the latter is sometimes considered with conformational sampling (see Consideration of Target Flexibility in Structure-Based Methods section). Furthermore, using structure-based methods, only targets with available structures can be queried, introducing a bias toward structurally known targets. Currently, most methods rely only on the available structures in the PDB. While there are over 140,000 protein structures deposited in the PDB (accessed in November 2018), they only cover at most 30% of the human proteome and 50% of known human drug targets,[126] with protein classes being differently well represented. Homology modeling is a possibility to infer lacking information from determined structures of homologous proteins. Somody et al.[126] have shown that given a sequence identity of ≥ 30% (as generally accepted lower limit for homology modeling) the structural coverage of the modeled human proteome could approach 70% (that of known human drug targets 95%). While large scale homology models have been used, e.g., for kinome-wide druggability predictions,[127] they have not been widely used yet for target prediction. It should be noted that the higher the sequence identity is, the more reliable the homology models are for structural modeling purposes. Furthermore, target-focused methods such as inverse screening and binding site comparison only require 3D target structures and binding site locations, whereas interactionfocused methods require ligand-target complex information, limiting their applicability. To overcome this, such interactions can be predicted: For instance, interaction fingerprint comparison can be coupled with inverse docking, and reverse pharmacophore screening can be based on target-focused pharmacophore methods such as T^2F-Pharm[128] that generate pharmacophores from apo-structures. However, it is important to note that such ligand- as well as structure-based models-based-on-models approaches may introduce noise to the predictions.

Target Ranking. Results from computational target prediction are highly dependent on the scoring function(s) used for target ranking. If two objects of the same type-for example, two small molecules or two protein binding sites- are compared, similarity of the query to the database can directly be inferred from the commonalities or mutual overlap between the objects and ranked accordingly. In contrast, if the objects to be compared are of different types, target ranking becomes more complex. For example, this is the case when the most likely targets are predicted for a small molecule based on individual machine learning models per target (ligand-based methods) or based on inverse screening against a target database (structure-based methods). While it is already challenging to predict the correct activity or binding energy of a ligand against one target, in panel predictions, the ligand is scored individually against multiple targets, requiring intertarget ranking. This is especially ambitious since the predictions are influenced by different forms of bias present in the data. Typically, some protein classes (e.g., kinases or G protein-coupled receptors) have been very well explored, whereas others have been explored less thoroughly (e.g., transporters). This means that more ligands are known for these proteins (ligand-based methods) or more structures have been elucidated (structure-based methods). Thus, the chemical or structural space is better covered, and they might score better

compared to less explored chemical or structural spaces. Another form of bias influencing target ranking can be the average molecular weight of ligands for certain protein classes. For example, the molecular weight of class B GPCRs is much higher than that of other proteins such as kinases. The higher molecular weight leads to the presence of more chemical substructures in the fingerprint vector and can increase the amount of predicted targets for these ligands.[58]

In an effort to reduce the effect of these biases on ligandbased prediction probability, raw probabilities can be converted to a z-score.[53] In this method, for all molecules in the training set, a prediction score is obtained for all proteins in the training set. Subsequently, for each protein, a mean probability and standard deviation of this probability can be derived and converted into a z-score. By applying the same z-scoring for novel compounds rather than the raw probability, the predictions are converted to a number of standard deviations over or under the mean for that particular protein. This method has been shown to be more robust than using the raw probability.[58] Similarly, in structure-based inverse screening, the interaction score of the ligand with each target is compared with the interaction score distribution from a set of reference ligands of the respective target complex structures, taken from X-ray structures or determined by docking.[81,83,84]

Validation Strategies. The performance of ligand-based models should always be estimated using external test sets to minimize overfitting (besides cross-validation). If test sets are composed randomly, this may lead to overoptimistic performance values as similar ligands may be present in both training and test sets, resulting in "easy" predictions. In order to overcome this effect, cluster splits, where the whole cluster of similar molecules is either contained in the test or training set, or temporal splits, where data from the most recent years is used for testing, can be applied.[129] Predictive performances of ligand-based models can be estimated by metrics such as R^2 and Q^2 as well as error-based metrics such as the root-meansquare error (RMSE) and mean absolute error (MAE). It is debatable what the best metric is to indicate model performance as this is dependent on the data and validation method. Generally, performance can be better estimated when multiple metrics are considered.[130]

Evaluating the performance of structure-based methods is based on diverse strategies. Binding site comparison methods, for instance, often screen a query target against a set of true (well-studied protein class with subclass classification) and decoy targets, whereas inverse screening methods often test only one or few query ligands in a set of true (known targets of the ligand) and decoy targets. Evaluation metrics are, for instance, the percentage of true targets in the top x% of the ranked hit list, the so-called enrichment factor (EF), and the area under the curve (AUC). While different sizes and compositions of benchmark data sets and the diverse use of performance metrics hamper a direct comparison between methods, efforts to unify benchmarking have been made. Since binding site comparison is a long-established approach with many published methods, proposed data sets have often been reused. Such an example is the data set compilation by Weill and Rognan,[99] encompassing a set of similar and dissimilar structure pairs as well as sets focused on kinases and serine endopeptidases (all scPDB-based). Also concentrating on similar and dissimilar pairs, Ehrt et al.[131] have recently proposed a collection of new and reused data sets (ProSPECCTs) to test different performance aspects, which the authors applied to multiple binding site methods to establish guidelines for their application scope. For inverse screening methods, Schomburg et al.[83] proposed two data sets

together with evaluation strategies: a small data set consisting of three target classes for detailed proof-of-concept and selectivity studies and a large data set with about 8000 protein structures and over 70 drug-like ligands. In addition to the widely used EF and AUC, the authors propose performance metrics capable of measuring the early enrichments, i.e., BEDROC (Boltzman-enhanced discrimination of ROC) and NSLR (normalized sum of logarithmic ranks).

Conclusion

Drug target identification is one of the most important, but also most complex, aspects of preclinical drug development. In this respect, computational target prediction is a highly valuable tool to identify the most probable targets for a compound under investigation. Such tools can guide wet lab experiments by suggesting potential targets for orphan compounds, supply tool compounds for functional analyses of poorly understood proteins, and thus help to decipher the mode-of-action of a protein under investigation. Furthermore, desired as well as undesired multitarget drug effects can be rationalized by computational (off-)target predictions, and known drugs can potentially be repositioned based on these forecasts.

Computational target prediction methods rely on the general assumption that similar molecules/structures will have similar interactions or interaction patterns. Exceptions are so-called activity cliffs, describing that small changes can cause large differences in activity.[29] Depending on the research question and the data available, ligand- or structure-based target prediction methods can be applied. In ligand-based methods, potential targets can either be inferred from the most similar known ligands or through elaborated machine learning models. The latter require sufficient and well annotated data in order to train proper models. Structure-based approaches compare a query protein based on their binding sites or interaction fingerprints to a panel of protein structures or screen a query compound against these panels using a docking or pharmacophore screening engine. It should be noted that usually ligand-centric methods are faster than structure-centric methods, especially when structural alignment or pose prediction is evoked. The former provides more quantitative information such as predicted bioactivities that can directly be associated with experimental values, whereas the latter can give additional information about the binding pose of ligands to potential targets. It should be noted that most methods do not consider alternate binding pockets on a single protein or the effect of protein complex formation. Although protein function or (de)activation through allosteric modulation can occur, most target prediction methods are based on the assumption that all ligands are orthosteric binders.

In our opinion, future progress needs to promote data coverage from both the ligand and protein point of view, e.g., annotation of non-biased bioactivities (reporting inactives) and deposition of novel structures or the same protein structures, but with different ligands to provide a better view on the dynamics of the ligand binding site (high-throughput crystallization). Furthermore, protein flexibility modeling and inter-target ranking are equally important matters to address. Moreover, new methods should be evaluated on standardized benchmarking data sets and performance metrics, as well as made accessible to the community in order to improve predictability, reliability, and reproducibility. Finally, holistic approaches should and will gain momentum, integrating multiple types of data, e.g., coupling chemical and structural space with information on the proteome level and pathways, linking cellular and molecular scales.

References

(1) Jenkins, J. L.; Bender, A.; Davies, J. W. In Silico Target Fishing: Predicting Biological Targets from Chemical Structure. Drug Discovery Today: Technol. 2006, 3, 413-421.
(2) Hart, C. P. Finding the Target After Screening the Phenotype. Drug Discovery Today 2005, 10, 513–519.
(3) Lee, J.; Bogyo, M. Target Deconvolution Techniques in Modern Phenotypic Profiling. Curr. Opin. Chem. Biol. 2013, 17, 118-126.
(4) Niphakis, M. J.; Cravatt, B. F. Enzyme Inhibitor Discovery by Activity-Based Protein Profiling. Annu. Rev. Biochem. 2014, 83, 341-377.
(5) Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. Mol. Inf. 2010, 29, 176-187.
(6) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. Pharmacol. Rev. 2014, 66, 334-395.
(7) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. Org. Biomol. Chem. 2004, 2, 3204.
(8) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. J. Med. Chem. 2014, 57, 787-7887.
(9) Morphy, R.; Kay, C.; Rankovic, Z. From Magic Bullets to Designed Multiple Ligands. Drug Discovery Today 2004, 9, 641-651.
(10) AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring Polypharmacology Using a ROCS-Based Target Fishing Approach. J. Chem. Inf. Model. 2012, 52, 492-505.
(11) Bender, A.; Scheiber, J.; Glick, M.; Davies, J.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. ChemMedChem 2007, 2, 861-873.
(12) Oprea, T. I.; et al. Drug Repurposing from an Academic Perspective. Drug Discovery Today: Ther. Strategies 2011, 8, 61-69.
(13) Keiser, M. J.; et al. Predicting New Molecular Targets for Known Drugs. Nature 2009, 462, 175-181.
(14) Ashburn, T. T.; Thor, K. B. Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. Nat. Rev. Drug Discovery 2004, 3, 673-683.
(15) Berger, A. B.; Vitorino, P. M.; Bogyo, M. Activity-Based Protein Profiling: Applications to Biomarker Discovery, in Vivo Imaging and Drug Discovery. Am. J. PharmacoGenomics 2004, 4, 371-381.
(16) Schirle, M.; Bantscheff, M.; Kuster, B. Mass Spectrometry- Based Proteomics in Preclinical Drug Discovery. Chem. Biol. 2012, 19, 72-84.
(17) van Esbroeck, A. C. M.; et al. Activity-Based Protein Profiling Reveals Off-Target Proteins of the FAAH Inhibitor BIa 10-2474. Science 2017, 356, 1084-1087.
(18) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100-D1107.
(19) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. Nucleic Acids Res. 2016, 44, D1202-D1213.
(20) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. J. Cheminf. 2017, 9, 17.
(21) Rognan, D. Chemogenomic Approaches to Rational Drug Design. Br. J. Pharmacol. 2007, 152, 38-52.
(22) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Model. 1988, 28, 31-36.
(23) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. Methods 2015, 71, 58-63.
(24) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742-754.
(25) Hawkins, P. C. D.; Stahl, G. In Computational Methods for GPCR Drug Discovery; Heifetz, A., Ed.; Springer: New York, 2018; pp 365-374.
(26) Shin, W.-H.; Zhu, X.; Bures, G. M.; Kihara, D. Three- Dimensional Compound Comparison Methods and Their Application in Drug Discovery. Molecules 2015, 20, 12841-62.
(27) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors
(MOLPRINT 2D): Evaluation of Performance. J. Chem. Inf. Comput. Sci. 2004, 44, 1708-1718.
(28) Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity. F1000Research 2013, 2, 144.
(29) Bajorath, J. Representation and Identification of Activity Cliffs. Expert Opin. Drug Discovery 2017, 12, 879-883.

(30) Subramanian, A.; et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 2017, 171, 1437-1452.
(31) De Wolf, H.; Cougnaud, L.; Van Hoorde, K.; De Bondt, A.; Wegner, J. K.; Ceulemans, H.; Göhlmann, H. High-Throughput Gene Expression Profiles to Define Drug Similarity and Predict Compound Activity. Assay Drug Dev. Technol. 2018, 16, 162-176.
(32) Simm, J.; et al. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chem. Biol. 2018, 25, 611-618.
(33) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data - a Statistical Analysis. PLoS One 2013, 8, No. e61007.
(34) Dalke, A. The FPS Fingerprint Format and Chemfp Toolkit. J. Cheminf. 2013, 5, P36.
(35) Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A Web Server for Target Prediction of Bioactive Small Molecules. Nucleic Acids Res. 2014, 42, W32-W38.
(36) Nickel, J.; Gohlke, B.-O.; Erehman, J.; Banerjee, P.; Rong, W. W.; Goede, A.; Dunkel, M.; Preissner, R. SuperPred: Update on Drug Classification and Target Prediction. Nucleic Acids Res. 2014, 42, W26-W31.
(37) Kauvar, L. M. Affinity Fingerprinting. Nat. Biotechnol. 1995, 13, 965-966.
(38) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining. J. Chem. Inf. Model. 2011, 51, 2440-2448.
(39) Lamb, J.; et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science 2006, 313, 1929-1935.
(40) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. Nat. Biotechnol. 2007, 25, 197.
(41) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. J. Mol. Biol. 1990, 215, 403-410.
(42) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 1997, 25, 3389-3402.
(43) Pearson, W. R. Empirical Statistical Estimates for Sequence Similarity Searches. J. Mol. Biol. 1998, 276, 71-84.
(44) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. Nature 2012, 486, 361.
(45) Czodrowski, P.; Bolick, W.-G. OCEAN: Optimized Cross REActivity EstimatioN. J. Chem. Inf. Model. 2016, 56, 2013-2023.
(46) Cortes Cabrera, A.; Lucena-Agell, D.; Redondo-Horcajo, M.; Barasoain, I.; Díaz, J. F.; Fasching, B.; Petrone, P. M. Aggregated Compound Biological Signatures Facilitate Phenotypic Drug Discovery and Target Elucidation. ACS Chem. Biol. 2016, 11, 3024-3034.
(47) Grover, A.; Grover, M.; Sharma, K. A Practical Overview of Quantitative Structure-Activity Relationship. World J. Pharm. Pharm. Sci. 2016, 5, 427-437.
(48) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 2003, 43, 1947-1958.
(49) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. Comput. Chem. 2001, 26, 5-14.
(50) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. J. Med. Chem. 2004, 47, 4463-4470.
(51) Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. Chem. Sci. 2018, 9, 513-530.
(52) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. Chemical Science 2019, 10, 1692.
(53) Mugumbate, G.; Abrahams, K. A.; Cox, J. A. G.; Papadatos, G.; van Westen, G.; Lelie vre, J.; Calus, S. T.; Loman, N. J.; Ballell, L.; Barros, D.; Overington, J. P.; Besra, G. S. Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and in Vitro Validation. PLoS One 2015, 10, No. e0121492.
(54) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. Bioinformatics 2000, 16, 747-748.
(55) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. J. Comput.-Aided Mol. Des. 2015, 29, 885-896.
(56) Afzal, A. M.; Mussa, H. Y.; Turner, R. E.; Bender, A.; Glen, R. C. A Multi-Label Approach to Target Prediction Taking Ligand Promiscuity into Account. J. Cheminf. 2015, 7, 24.
(57) Anger, L. T.; Wolf, A.; Schleifer, K.-J.; Schrenk, D.; Rohrer, S. G. Generalized Workflow for Generating Highly Predictive in Silico Off-Target Activity Models. J. Chem. Inf. Model. 2014, 54, 2411-2422.
(58) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.;

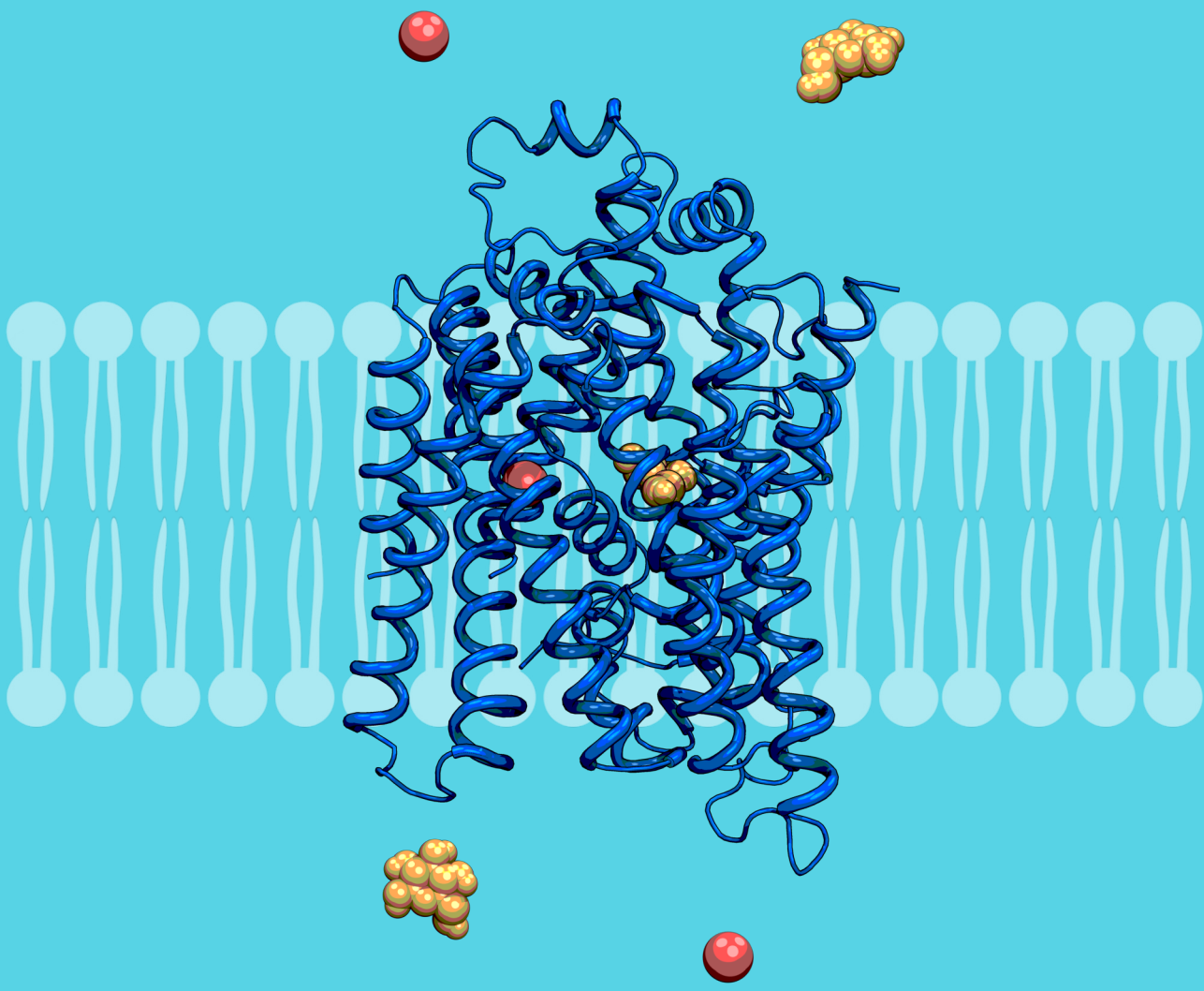
van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. J. Cheminf. 2017, 9, 45.
(59) Zhang, Q.; Hughes-Oliver, J. M.; Ng, R. T. A Model-Based Ensembling Approach for Developing QSARs. J. Chem. Inf. Model. 2009, 49, 1857-1865.
(60) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. MedChemComm 2011, 2, 16-30.
(61) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. a Multivariate Characterization of 87 Amino Acids. J. Med. Chem. 1998, 41, 2481-2491.
(62) van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. J. Cheminf. 2013, 5, 42.
(63) van Westen, G. J. P.; van den Hoven, O. O.; van der Pijl, R.; Mulder-Krieger, T.; de Vries, H.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. J. Med. Chem. 2012, 55, 7010-7020.
(64) van Westen, G. J. P.; Wegner, J. K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. PLoS One 2011, 6, No. e27518.
(65) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large- Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. Chem. Sci. 2018, 9, 5441-5451.
(66) Oprea, T. I.; Nielsen, S. K.; Ursu, O.; Yang, J. J.; Taboureau, O.; Mathias, S. L.; Kouskoumvekaki, l.; Sklar, L. A.; Bologa, C. G. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. Mol. Inf. 2011, 30, 100-111.
(67) Lo, Y.-C.; Senese, S.; Damoiseaux, R.; Torres, J. Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. ACS Chem. Biol. 2016, 11, 2244-2253.
(68) Yamanishi, Y.; Kotera, M.; Moriya, Y.; Sawada, R.; Kanehisa, M.; Goto, S. DINIES: Drug-Target Interaction Network Inference Engine Based on Supervised Analysis. Nucleic Acids Res. 2014, 42, W39-W45.
(69) Zhao, S.; Li, S. Network-Based Relating Pharmacological and Genomic Spaces for Drug Target Identification. PLoS One 2010, 5, No. e11764.
(70) Chen, X.; Liu, M.-X.; Yan, G.-Y. Drug-Target Interaction Prediction by Random Walk on the Heterogeneous Network. Mol. BioSyst. 2012, 8, 1970-1978.
(71) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of Drug-target Interaction Networks from the Integration of Chemical and Genomic Spaces. Bioinformatics 2008, 24, i232-i240.
(72) OpenEye Scientific Software, ROCS. https://www.eyesopen. com/rocs
(accessed on 2018-11-01).
(73) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. J. Comput.-Aided Mol. Des. 1998, 12, 471-490.
(74) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. J. Cheminf. 2015, 7, 51.
(75) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Res. 2000, 28, 235-242.
(76) Volkamer, A.; von Behren, M. M.; Bietz, S.; Rarey, M. Applied Chemoinformatics; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2018; pp 283-311.
(77) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-Database of Ligandable Binding Sites-10 Years On. Nucleic Acids Res. 2015, 43, D399-D404.
(78) Xu, X.; Huang, M.; Zou, X. Docking-Based Inverse Virtual Screening: Methods, Applications, and Challenges. Biophys. Rep. 2018, 4, 1-16.
(79) Huang, H.; Zhang, G.; Zhou, Y.; Lin, C.; Chen, S.; Lin, Y.; Mai, S.; Huang, Z. Reverse Screening Methods to Search for the Protein Targets of Chemopreventive Compounds. Front. Chem. 2018, 6, 138.
(80) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: A Web Server for Identifying Drug Targets with Docking Approach. Nucleic Acids Res. 2006, 34, W219-W224.
(81) Wang, J.-C.; Chu, P.-Y.; Chen, C.-M.; Lin, J.-H. idTarget: A Web Server for Identifying Protein Targets of Small Chemical Molecules with Robust Scoring Functions and a Divide-And-Conquer Docking Approach. Nucleic Acids Res. 2012, 40, W393-W399.
(82) Chen, Y. Z.; Zhi, D. G. Ligand-Protein Inverse Docking and Its Potential Use in the Computer Search of Protein Targets of a Small Molecule. Proteins: Struct., Funct., Genet. 2001, 43, 217-226.
(83) Schomburg, K. T.; Bietz, S.; Briem, H.; Henzler, A. M.; Urbaczek, S.; Rarey, M. Facing the Challenges of Structure-

Based Target Prediction by Inverse Virtual Screening. J. Chem. Inf. Model. 2014, 54, 1676-1686.
(84) Santiago, D. N.; Pevzner, Y.; Durand, A. A.; Tran, M.; Scheerer, R. R.; Daniel, K.; Sung, S.-S.; Lee Woodcock, H.; Guida, W. C.; Brooks, W. H. Virtual Target Screening: Validation Using Kinase Inhibitors. J. Chem. Inf. Model. 2012, 52, 2192-2203.
(85) Ellingson, S. R.; Smith, J. C.; Baudry, J. VinaMPI: Facilitating Multiple Receptor High-Throughput Virtual Docking on High- Performance Computers. J. Comput. Chem. 2013, 34, 2212-2221.
(86) Li, G.-B.; Yu, Z.-J.; Liu, S.; Huang, L.-Y.; Yang, L.-L.; Lohans, C. T.; Yang, S.-Y. IFPTarget: A Customized Virtual Target Identification Method Based on Protein-Ligand Interaction Fingerprinting Analyses. J. Chem. Inf. Model. 2017, 57, 1640-1651.
(87) Scafuri, B.; Marabotti, A.; Carbone, V.; Minasi, P.; Dotolo, S.; Facchiano, A. A Theoretical Study on Predicted Protein Targets of Apple Polyphenols and Possible Mechanisms of Chemoprevention in Colorectal Cancer. Sci. Rep. 2016, 6, 32516.
(88) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). Pure Appl. Chem. 1998, 70, 1129-1143.
(89) Schuster, D. 3D Pharmacophores As Tools for Activity Profiling. Drug Discovery Today: Technol. 2010, 7, e205-e211.
(90) Wang, X.; Shen, Y.; Wang, S.; Li, S.; Zhang, W.; Liu, X.; Lai, L.; Pei, J.; Li, H. PharmMapper 2017 Update: A Web Server for Potential Drug Target Identification with a Comprehensive Target Pharmacophore Database. Nucleic Acids Res. 2017, 45, W356-W360.
(91) Liu, B.; Fu, X.-Q.; Li, T.; Su, T.; Guo, H.; Zhu, P.-L.; Tse, A. K.-W.; Liu, S.-M.; Yu, Z.-L. Computational and Experimental Prediction of Molecules Involved in the Anti-Melanoma Action of Berberine. J. Ethnopharmacol. 2017, 208, 225-235.
(92) Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? Curr. Comput.-Aided Drug Des. 2008, 4, 209-220.
(93) Ehrt, C.; Brinkjost, T.; Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. J. Med. Chem. 2016, 59, 4121-4151.
(94) Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. J. Mol. Biol. 2002, 323, 387-406.
(95) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. Proteins: Struct., Funct., Genet. 2003, 52, 137-145.
(96) Moriaud, F.; Richard, S. B.; Adcock, S. A.; Chanas-Martin, L.; Surgand, J.-S.; Ben Jelloul, M.; Delfaud, F. Identify Drug Repurposing Candidates by Mining the Protein Data Bank. Briefings Bioinf. 2011, 12, 336-340.
(97) Konc, J.; Janezic, D. ProBiS Algorithm for Detection of Structurally Similar Protein Binding Sites by Local Structural Alignment. Bioinformatics 2010, 26, 1160-1168.
(98) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. J. Chem. Inf. Model. 2012, 52, 2287-2299.
(99) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. J. Chem. Inf. Model. 2010, 50, 123-135.
(100) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-Throughput Virtual Screening of Proteins Using GRID Molecular Interaction Fields. J. Chem. Inf. Model. 2010, 50, 155-169.
(101) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. J. Mol. Biol. 2004, 339, 607-633.
(102) von Behren, M. M.; Volkamer, A.; Henzler, A. M.; Schomburg, K. T.; Urbaczek, S.; Rarey, M. Fast Protein Binding Site Comparison Via an Index-Based Screening Technology. J. Chem. Inf. Model. 2013, 53, 411-422.
(103) Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large- Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real. J. Chem. Inf. Model. 2015, 55, 165-179.
(104) Brakoulias, A.; Jackson, R. M. Towards a Structural Classification of Phosphate Binding Sites in Protein-Nucleotide Complexes: An Automated All-Against-All Structural Comparison Using Geometric Matching. Proteins: Struct., Funct., Genet. 2004, 56, 250-260.
(105) Aung, Z.; Tong, J. C. BSAlign: A Rapid Graph-Based Algorithm for Detecting Ligand-Binding Sites in Protein Structures. Genome Inform 2008, 21, 65-76.
(106) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand- Binding Sites. Proteins: Struct., Funct., Genet. 2008, 71, 1755-1778.
(107) De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of Protein Kinase Inhibitors to Synapsin I Inferred from Pair-Wise Binding Site Similarity Measurements. PLoS One 2010, 5, No. e12214.
(108) Salentin, S.; Haupt, V. J.; Daminelli, S.; Schroeder, M. Polypharmacology Rescored: Protein-Ligand Interaction Profiles for Remote Binding Site Similarity Assessment. Prog. Biophys. Mol. Biol. 2014, 116, 174-186.

(109) Cao, R.; Wang, Y. Predicting Molecular Targets for Small- Molecule Drugs with a Ligand-Based Interaction Fingerprint Approach. ChemMedChem 2016, 11, 1352-1361.
(110) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. J. Med. Chem. 2004, 47, 337-344.
(111) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. J. Chem. Inf. Model. 2013, 53, 623-637.
(112) Da, C.; Kireev, D. Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. J. Chem. Inf. Model. 2014, 54, 2555-2561.
(113) Kinoshita, K.; Nakamura, H. Identification of Protein Biochemical Functions by Similarity Search Using the Molecular Surface Database EF-Site. Protein Sci. 2003, 12, 1589-1595.
(114) Reisen, F.; Weisel, M.; Kriegl, J. M.; Schneider, G. Self- Organizing Fuzzy Graphs for Structure-Based Comparison of Protein Pockets. J. Proteome Res. 2010, 9, 6498-6510.
(115) Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. BMC Bioinf. 2008, 9, 543.
(116) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. J. Chem. Inf. Model. 2012, 52, 2031-2043.
(117) Liu, T.; Altman, R. B. Using Multiple Microenvironments to Find Similar Ligand-Binding Sites: Application to Kinase Inhibitor Binding. PLoS Comput. Biol. 2011, 7, No. e1002326.
(118) Siragusa, L.; Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. BioGPS: Navigating Biological Space to Predict Polypharmacology, Off-Targeting, and Selectivity. Proteins: Struct., Funct., Genet. 2015, 83, 517-532.
(119) Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C. TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. J. Chem. Inf. Model. 2013, 53, 1235-1252.
(120) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. Nucleic Acids Res. 2006, 34, D668-D672.
(121) Hilbert, M.; López, P. The World's Technological Capacity to Store, Communicate, and Compute Information. Science 2011, 332, 60-65.
(122) Eeckhout, L. Is Moore's Law Slowing Down? What's Next? IEEE Micro 2017, 37, 4-5.
(123) Jasial, S.; Hu, Y.; Bajorath, J. Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping. J. Chem. Inf. Model. 2016, 56, 300-307.
(124) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? J. Chem. Inf. Model. 2006, 46, 2445-2456.
(125) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. J. Chem. Inf. Model. 2013, 53, 2499-2505.
(126) Somody, J. C.; MacKinnon, S. S.; Windemuth, A. Structural Coverage of the Proteome for Pharmaceutical Applications. Drug Discovery Today 2017, 22, 1792-1799.
(127) Volkamer, A.; Eid, S.; Turk, S.; Jaeger, S.; Rippmann, F.; Fulle, S. Pocketome of Human Kinases: Prioritizing the ATP Binding Sites of (Yet) Untapped Protein Kinases for Drug Discovery. J. Chem. Inf. Model. 2015, 55, 538-549.
(128) Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. Molecules 2018, 23, 1959.
(129) Sheridan, R. P. Time-Split Cross-Validation As a Method for Estimating the Goodness of Prospective Prediction. J. Chem. Inf. Model. 2013, 53, 783-790.
(130) Gramatica, P.; Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. J. Chem. Inf. Model. 2016, 56, 1127-1131.
(131) Ehrt, C.; Brinkjost, T.; Koch, O. A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor- Made Data Sets.

Chapter three

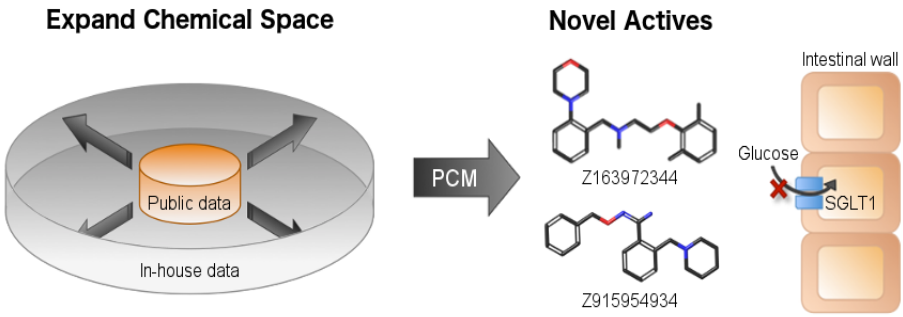
Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 using Proteochemometric Modeling



Lindsey Burggraaff, Paul Oranje, Robin Gouka, Pieter van der Pijl, Marian Geldof, Herman W.T. van Vlijmen, Adriaan P. IJzerman, and Gerard J.P. van Westen. Journal of Cheminformatics. 2019. doi: 10.1186/s13321-019-0337-8.

Abstract

Sodium-dependent glucose co-transporter 1 (SGLT1) is a solute carrier responsible for active glucose absorption. SGLT1 is present in both the renal tubules and small intestine. In contrast, the closely related sodium-dependent glucose co-transporter 2 (SGLT2), a protein that is targeted in the treatment of diabetes type II, is only expressed in the renal tubules. Although dual inhibitors for both SGLT1 and SGLT2 have been developed, no drugs on the market are targeted at decreasing dietary glucose uptake by SGLT1 in the gastrointestinal tract. Here we aim at identifying SGLT1 inhibitors *in silico* by applying a machine learning approach that does not require structural information, which is absent for SGLT1. We applied proteochemometrics by implementation of compound- and protein-based information into random forest models. We obtained a predictive model with a sensitivity of 0.64 ± 0.06, specificity of 0.93 ± 0.01, positive predictive value of 0.47 ± 0.07, negative predictive value of 0.96 ± 0.01, and Matthews correlation coefficient of 0.49 ± 0.05. Subsequent to model training, we applied our model in virtual screening to identify novel SGLT1 inhibitors. Of the 77 tested compounds, 30 were experimentally confirmed for SGLT1-inhibiting activity in vitro, leading to a hit rate of 39% with activities in the low micromolar range. Moreover, the hit compounds included novel molecules, which is reflected by the low similarity of these compounds with the training set (< 0.3). Conclusively, proteochemometric modeling of SGLT1 is a viable strategy for identifying active small molecules. Therefore, this method may also be applied in detection of novel small molecules for other transporter proteins.



Introduction

Sodium-dependent glucose co-transporters, or sodium-glucose linked transporters (SGLTs), are solute carriers (SLCs) that are responsible for glucose (re)absorption. SGLTs are members of the sodium-dependent transporters and are encoded by the SLC5A genes.[1] SGLTs are interesting targets in the treatment of diabetes mellitus, as their inhibition reduces the risk of hyperglycemia by decreasing glucose (re-)uptake.[2] In the human body two SGLT isoforms are involved in glucose transport: SGLT1 and SGLT2.[3] Both SGLT1 and SGLT2 are expressed in the kidney, whereas SGLT1 is also expressed in the small intestine.[4] SGLT2 is a high capacity transporter responsible for 90% of glucose reuptake in the renal tubules and multiple compounds have been developed that inhibit this solute carrier.[5,6] Furthermore, SGLT2 inhibition has been shown to decrease blood glucose levels in diabetes type 2 patients.[7] In contrast to SGLT2, SGLT1 is a low-capacity glucose transporter.[1] However, SGLT1 has a higher glucose affinity than SGLT2 and is additionally capable of transporting galactose.[1] Dual inhibitors blocking both SGLT1 and SGLT2 are currently in clinical development.[8,9] In line with previous evidence we suggest that SGLT1 inhibition in the intestine will lower blood glucose levels as well.[10,11] Compounds that do not penetrate the intestinal wall can achieve selective targeting of SGLT1 in the intestine, as they would not reach the renal tubules.[12]

The complexity and the hydrophobic nature of transporter proteins make them challenging to crystalize. Crystal structures of transporters are scarce and binding locations of small molecules to these transporters are often unknown. For human SGLTs no protein structures are available negating the use of structure-based modeling techniques. However, the publicly available compound database ChEMBL includes ligand-protein binding information for multiple SGLTs,[13-15] allowing the use of statistical modeling techniques such as quantitative structure-activity relationship analysis (QSAR) and proteochemometrics (PCM).[16] These techniques, which make use of machine learning, do not require protein structural information and can therefore more easily be applied in the context of SLCs. Although ligand-based pharmacophore modeling, QSAR, and PCM have only been applied to a few SLCs,[17,18] these techniques are well established on other drug targets including membrane proteins such as G protein-coupled receptors.[19-21]

Unfortunately, the publicly available compound interaction data for SGLTs is limited from the point of chemical diversity as the major share of ligands are glycoside-like compounds and oxopyrrolidine-carboxamides. This limited chemical space hence restricts the applicability domain of QSAR and PCM models.[22] The applicability domain of computational models can be interpreted as the theoretical ensemble of molecular structures to which a model can be applied accurately. This domain is dependent on the model input and can therefore be quantified by similarity with the training molecules.

In the current work we show how we expanded the chemical space of SGLT inhibitors (using an in-house dataset),[23] and with that the applicability domain of our SGLT models. We constructed PCM models based on SGLT1 and its closest family members to predict compound activity for SGLT1. We successfully identified novel SGLT1 inhibitors that display low similarity towards the training set.

Results and Discussion

SGLT chemical space

A public dataset was created based on ChEMBL version 23[13,15] which includes the target protein human SGLT1 (hSGLT1), related protein human SGLT2 (hSGLT2), and multiple other SGLTs from different species. The public dataset encompassed 2,063 data points and 1,683 unique compounds, of which 886 compounds had measured hSGLT1 activities. Additionally, this set was supplemented with an in-house dataset of 2,007 molecules previously screened for hSGLT1 and hSGLT2 inhibition.[23] This in-house dataset has as its basis the Spectrum Collection compound library[24] extended with compounds similar to primary screening hits and contained natural products and synthetic compounds. The data derived from ChEMBL was compared to the in-house dataset: the in-house dataset contained an additional 2,005 hSGLT1 activities and 140 hSGLT2 activities, which were not present in the public dataset. The difference between the public and in-house dataset is graphically represented with t-Distributed Stochastic Neighbor Embedding (t-SNE)[25] (**Figure 3.1A**). This technique was applied to decrease the high dimensionality of the datasets, making it possible to visualize them. The high dimensions are a consequence of the many descriptors that are used to describe the data, i.e. FCFP6 fingerprints. The t-SNE plot shows that the data derived for proteins similar to hSGLT1 extend the chemical space; many hSGLT2 compounds from the public domain are not tested on hSGLT1 and thus provide additional chemical information. The in-house and public datasets considerably differ from each other, with a slight overlap of only a few hSGLT1 and hSGLT2 public compounds with the in-house dataset. To further investigate the difference between the public and in-house dataset, the following physicochemical properties were considered: molecular weight, ALogP, and number of hydrogen bond donors and acceptors. The publicly available data represented mainly the drug-like space, following Lipinski's rule of five, likely resulting from the fact that hSGLT2 is a drug target investigated by pharmaceutical companies.[26] Moreover, the public data mostly includes glycoside-like compounds and oxopyrrolidine-carboxamides. In contrast, the in-house dataset encompasses more diverse molecules and captures a wider value range for the physicochemical properties mentioned above. The molecular weight and ALogP are represented in **Figure 3.1B**, where it is observed that these properties are more conserved for the public dataset than for the in-house dataset. Additionally, the number of hydrogen bond donors and acceptors is more widely captured by the in-house dataset (mean and standard deviation): public dataset hydrogen bond donor 3.6 ±1.6, hydrogen bond acceptor 6.3 ±1.8; in-house dataset hydrogen bond donor 2.0 ±2.6, hydrogen bond acceptor 5.1 ±4.1. When screening for compounds to target hSGLT1 in the intestine, it is favorable to consider compounds that do not necessarily adhere to Lipinski's rule of five, as it is preferred to minimize compound absorption from the gastrointestinal tract. Therefore, the in-house dataset contributes substantially to the applicability domain and predictive power of the statistical SGLT model.

Merging different datasets

To merge the public and in-house dataset the difference in activity units for both sets had to be resolved. The public dataset contains pChEMBL values, representing a standardized unit for affinity and potency values such as K_i, IC_{50}, EC_{50}, and K_d.[27] The potency values in the in-house dataset were available as percentage activity compared to (negative) control at a concentration of 50 μM, which could not be converted into a pChEMBL value. Hence, binary classification models were chosen over regression.

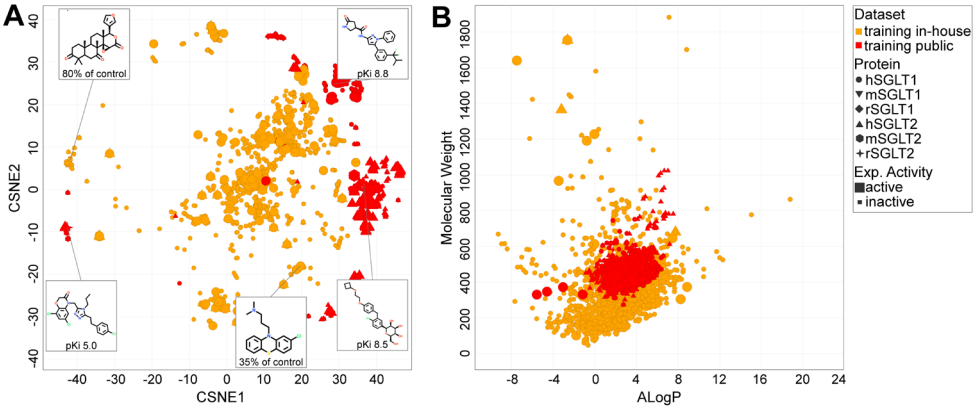


Figure 3.1. Chemical space of the public and in-house datasets. A) the t-SNE shows molecular structure and affinity (pK_i for public data and % of (negative) control for in-house data) for representative hSGLT1 compounds. B) molecular weight and ALogP distribution of compounds in the training sets.

Thresholds for compounds being 'active' were determined by grid searching cut-off values for both the public and in-house data. Activity thresholds along the grid were reviewed using hSGLT1 QSARs and external validation with a hold-out test set containing 30% of the in-house hSGLT1 data. The public domain compounds, which are mostly glycoside-like compounds and oxopyrrolidine-carboxamides, only describe a very conserved and small chemical space. However, the molecules of interest belonged to the same chemical space as the more diverse in-house compounds and therefore only compounds from the in-house set were used in validation. The activity threshold grid search showed that an activity threshold optimum for the in-house data was found at activity percentage < 70%, < 75%, and < 80% together with the threshold for public data set at pChEMBL > 8.5 (**Figure 3.2**). In further models the activity threshold was set at activity < 70% for in-house data and pChEMBL > 8.5 for public data to achieve the best performance for predicting hSGLT1 active molecules in the chemical space of the in-house compounds. Although these activity thresholds are not similar toward each other (e.g. pChEMBL > 8.5 corresponds to an in-house threshold much lower than 70%), these thresholds were determined optimal for the aim, which is the identification of novel (weak) actives that are similar in chemical space as the in-house compounds. The performance of the QSAR benchmark model using the selected thresholds was: sensitivity 0.76, specificity 0.86, positive predictive value (PPV) 0.42, negative predictive value (NPV) 0.96, and Matthews correlation coefficient (MCC) 0.48.

Proteochemometric modeling of hSGLT1

A PCM model was constructed using only public data to predict the inhibitory activity of compounds for hSGLT1. The performance of the model was tested on in-house data as these compounds represented the chemical space of interest. The model was validated using five test sets composed from in-house hSGLT1 data (5x20%). The mean performance of the public data model was very poor (mean with standard deviation): sensitivity 0.01 ±0.01, specificity 0.98 ±0.00, PPV 0.03 ±0.06, NPV 0.91 ±0.01, and MCC -0.03 ±0.03 (**Table 3.1**). Apparently, with the public data alone it was impossible to identify active compounds and the process defaulted to classification of all compounds as 'inactive'. This behavior confirms the large differences in

chemical space between the two sets as alluded to above.

Next, a PCM model was constructed based on the combined full data set consisting of all public and in-house data. To validate the performance of this model, 5-fold cross-validation was applied with the same test sets as applied in validation of performance of the public data model: rotationally 20% of the in-house hSGLT1 data was used as holdout test set; the remaining 80% was used in training. In each case the test set contained compounds not available for training. This resulted in the following performance: sensitivity 0.64 ±0.06, specificity 0.93 ±0.01, PPV 0.47 ±0.07, NPV 0.96 ±0.01, and MCC 0.49 ±0.05. Overall performance of this PCM model was regarded satisfactory for predictions of new compounds and was comparable with the QSAR benchmark model used for activity threshold determination.

Additionally the performance of models trained on in-house data only was tested to assess the effect of addition of public data. Public domain compounds contributed slightly to the predictive performance of the model in specificity, PPV, and MCC. This was observed by a minor decrease in performance upon removal of the public data from the training set: sensitivity 0.69 ±0.07, specificity 0.89 ±0.02, PPV 0.38 ±0.06, NPV 0.97 ±0.01, and MCC 0.45 ±0.05. Although the difference in performances is not significant, it is remarkable that the number of false positives decreases considerably when public data is included in training, whereas the number of true positives is only slightly negatively affected: false positives 28 ±6 vs. 43 ±6, true positives 24 ±4 vs. 26 ±4 (with and without public data, respectively). Apparently, the public data by itself is not sufficient in predicting hSGLT1 activity in the chemical space of the in-house compounds but does add value to the model when supplemented to the in-house dataset.

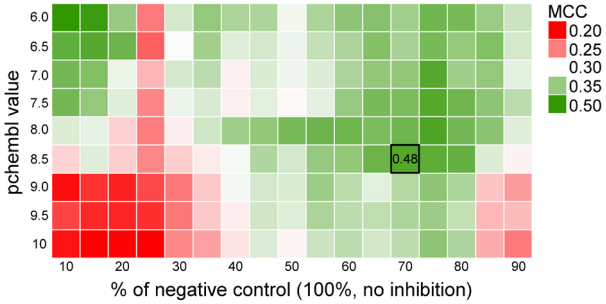


Figure 3.2. Activity threshold grid search. Searching the activity threshold grid for in-house (activity percentage compared to negative control) and public data (pChEMBL value). Model performance was measured using Matthews Correlation Coefficient (MCC), which was 0.48 for the final selected thresholds of < 70% for in-house data and pChEMBL > 8.5 for public data.

Table 3.1. Model performance depends on datasets that are used in training.

Model and validation	Training	Sensitivity	Specificity	PPV	NPV	MCC
QSAR (EV)	PD + IH	0.76	0.86	0.42	0.96	0.48
Public PCM (CV)	PD	0.01 ±0.01	0.98 ±0.00	0.03 ±0.06	0.91 ±0.01	-0.03 ±0.03
In-house PCM (CV)	IH	0.69 ±0.07	0.89 ±0.02	0.38 ±0.06	0.97 ±0.01	0.45 ±0.05
Combined PCM (CV)	PD + IH	0.64 ±0.06	0.93 ±0.01	0.47 ±0.07	0.96 ±0.01	0.49 ±0.05

Abbreviations: PD - public data, IH - in-house data, EV - external validation on 30% of data, CV - 5-fold cross validation on 20% of the data per iteration.

Screening for hSGLT1 actives in a commercially available compound library

The SGLT PCM model that was trained on public and in-house data was applied to a commercially available library. This library, the Enamine high-throughput screening (HTS) library, contains over 1.8 million compounds.[28] The library covers a wide diversity regarding molecular weight and ALogP values, and encompasses a vast chemical space (**Figure 3.3**). With the PCM model, an hSGLT1 activity prediction was assigned to all 1,815,674 compounds in the library (model training time was 103 seconds; the screening speed was approximately 132 seconds for 10,000 compounds). 155,275 compounds were predicted to be in the active class based on a predicted class probability of ≥ 0.5 (score, average of the trees in agreement with the active class).

To increase confidence in the activity of compounds the screened set was pre-filtered by selecting compounds with a predicted class probability of ≥ 0.8 on a scale from 0 to 1. Here, a resulting score of 1 represents compounds predicted to be in the 'active' class, a score of 0 indicates that the compounds are predicted 'inactive'; ascending scores indicate higher certainty of compounds belonging to the 'active' class. Additionally, compounds with molecular weight ≤ 300 were removed to exclude fragment-like compounds. The final filtered set contained 672 compounds.

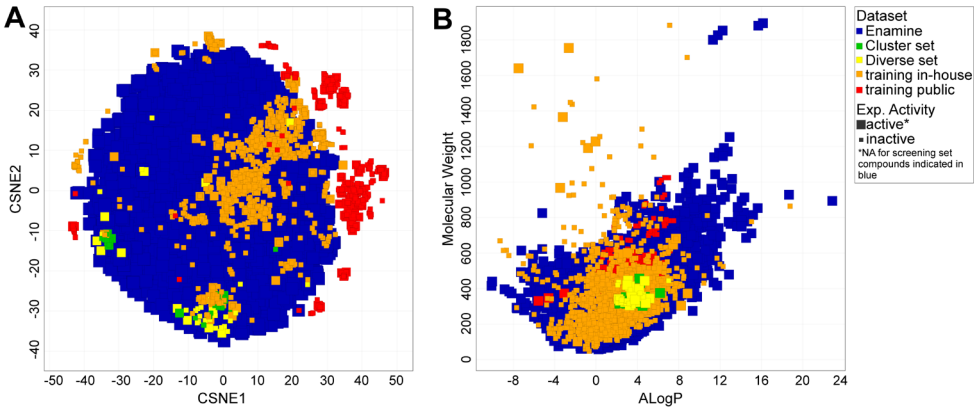


Figure 3.3. Chemical space of the selected compounds compared to the training and screening datasets. A) the Diverse set (yellow) and Cluster set (green) are displayed compared to the training (orange and red) and Enamine screening set (blue). The Enamine set is represented by a random selection of 20,000 out of the total of 1,815,674 compounds (~1%) in the screening set to limit t-SNE calculation time. B) the molecular weight and ALogP of the Diverse and Cluster set compared to the training and screening sets.

Based on the model predictions, 40 chemically diverse compounds predicted to be active were selected for experimental in vitro validation ('Diverse set'). The compounds in this set were cluster centers resulting from clustering of the remaining predicted active compounds into 40 clusters. This Diverse set was selected to increase the probability of detecting chemically novel hSGLT1 inhibitors. The selected compounds distributed widely through chemical space (**Figure 3.3** and **Supplementary Figure S3.1**), thus providing a challenging test for the SGLT PCM model. In addition to screening for novel hSGLT1 inhibitors, compounds were selected to expand the SAR around some recently identified hSGLT1 inhibitors from the in-house dataset.[23] Based on four hSGLT1 inhibitors (**Figure 3.4**) 3x10 additional compounds were selected from the pre-filtered Enamine HTS set that were predicted to be active (with top ranking scores) and that

resembled bepridil, bupivacaine, and cloperastine and 7 compounds were selected resembling trihexyphenidyl ('Cluster set'). These compounds were selected based on both model prediction (predicted class probability ≥ 0.8) and the highest similarity (Tanimoto, FCFP6) towards their known reference compound.

The total selection of 77 unique compounds was tested in vitro in cells expressing hSGLT1 in a single point measurement at a concentration of 50 μM. From the 40 diverse predicted hits that were assessed, 15 compounds were defined active as they displayed hSGLT1 inhibition in vitro with an activity reaching values below 70% compared to the negative control (100%, no inhibition) (**Supplementary Table S3.2**). From the 37 Cluster set compounds, an additional 15 compounds were confirmed to be active (**Supplementary Table S3.3**).

Cytotoxicity of hSGLT1 actives

The potential cytotoxicity of the screening compounds (Diverse set and Cluster set) was investigated by analysis of secreted adenylate kinase (AK), a marker of cell wall integrity loss. Most compounds did not show any indication of cyotoxicity, however one active from the Diverse set displayed moderate impairment of the cell wall (Z1416510792: activity 43 ±9%, cytotoxicity 25%). The cytotoxicity assay was limited by the available supernatant from the activity screen. Therefore not all compounds were measured in duplicate and cytotoxicity of one active from the Cluster set could not be determined (Z817504494: activity 45 ±3%).

Compound activity for hSGLT2

Both the Diverse set and Cluster set compounds were additionally measured for hSGLT2 inhibitory activity to assess their selectivity between the two transporters. The same cellular screening assay was performed as was used for hSGLT1 (single point measurement at a concentration of 50 μM). More actives were defined for hSGLT2 compared to hSGLT1 using the same activity threshold of 70% activity relative to negative control (100%, no inhibition): 22 actives in the Diverse set and 19 in the Cluster set. Almost all hSGLT1 actives showed activity for hSGLT2 with the possible exception of Z105569118, which only marginally surpassed the activity threshold for hSGLT2 (activity of hSGLT1 64 ±4% and hSGLT2 76 ±5%). No selective compounds were identified for hSGLT1, with 14% being the highest observed difference in inhibition (Z46160496: hSGLT1 41 ±4% and hSGLT2 55 ±2%). For hSGLT2 the biggest difference in inhibition was found for Z1318177320 that showed a difference of 39% (hSGLT1 93 ±20% and hSGLT2 54 ±0%).

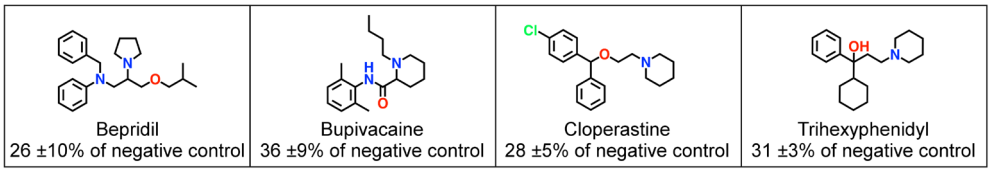


Figure 3.4. Reference hSGLT1 inhibitors for Cluster set and their inhibitory activity. Inhibitory activities (compared to negative control, where 100% is no inhibition) and chemical structures of four recently identified novel hSGLT1 inhibitors: bepridil, bupivacaine, cloperastine, and trihexyphenidyl.

Hit compound analysis

The activities of the hit compounds of the Diverse and Cluster set were analyzed. The strongest inhibitors, Z163972344 and Z915954934, were derived from the Diverse set with activities of 24 ±1% and 28 ±4% (100%: no inhibition), respectively. Z163972344 has low similarity (0.27 based on Tanimoto FCFP6) with the training set, indicating that this is a truly novel inhibitor for hSGLT1. The average similarity of actives in the Diverse set compared to training was 0.33, with Z1416510792 being the active that is most similar to the compounds in the training set with a similarity score of 0.61 (this compound showed moderate AK secretion in the cytotoxicity assay).

For the Cluster set a total of 15 actives were validated for the four different clusters. The cloperastine cluster encompassed the most actives (60% actives), whereas the trihexyphenidyl and bepridil clusters contained the least actives with 29% and 30% actives, respectively. The bupivacaine cluster had a hit rate of 40%, which is comparable with the overall hit rate of the total Cluster set (41%). The variance in hit rates between the four clusters is also reflected in the similarity of compounds toward their cluster reference: the cloperastine and bupivacaine clusters contained the most similar compounds (average similarities towards cluster reference compound were 0.43 and 0.42, respectively); the trihexyphenidyl and bepridil clusters contained less similar compounds (0.35 and 0.31, respectively).

Although the cloperastine and bupivacaine clusters contained the most similar cluster members, no conclusive SAR could be determined. The cluster members displayed variations in methyl substituents, which showed an effect for two compounds in the bupivacaine cluster (Z46224544 (45 ±10%) and Z2217101732 (74 ±8%)). This was however not observed for compounds in the cloperastine cluster: Z31367782 (36 ±4%), Z31371621 (37 ±3%), Z31367784 (43 ±7%), and Z31370217 (45 ±10%). The positions of the methyl substituents were too distinct to make solid conclusions on their relationship with compound activity.

In general, the novel active entities contain at least one aromatic ring and two hydrogen bond acceptors. Only two of the 30 actives did not adhere to Lipinski's rule of five, with an ALogP of 5.2 and 6.2 for Z1844922248 (activity 49 ±7%) and Z56906862 (activity 38 ±5%), respectively.

Aiming for specific targeting at the gastrointestinal tract

AAs mentioned in the Introduction, hSGLT1 inhibition at the intestinal wall is desired. Based on chemical structure and physicochemical properties the identified hit compounds will most likely be absorbed. However, it is suggested that modifications can be introduced to improve specific intestinal targeting. These alterations, such as a higher molecular weight, can prevent compounds from being absorbed or transported by the intestinal wall.[29] Intestinal SGLT1 blockers are expected to display less renal damage, which is an adverse effect observed for SGLT2 inhibitors.[6] Moreover, drug action restricted to the gastrointestinal tract also limits other off-target interactions, which were observed for the marketed SGLT2 inhibitor canagliflozin.[30] An example of a compound that was optimized for specific targeting at the gastrointestinal tract is LX2761, an inhibitor aimed at intestinal SGLT1 that decreased glucose uptake in mice.[31,32] Although SGLT1 inhibition at the intestine may not comprise renal function, other adverse effects that can result from intestinal targeting need to be considered.[33,34]

Indications for alternate binding modes

Upon examination of our hSGLT1 actives, a large variety in chemical structure and physicochemical properties was observed. This indicates that different ligand types may bind to different sites on hSGLT1. It is speculated that the glycoside-like hSGLT1 inhibitors, which are represented well in the public compound domain, bind to the glucose binding site, whereas more chemically diverse hSGLT1 inhibitors are suggested to bind either there or elsewhere on the protein. The hSGLT1 actives were grouped into ten clusters. Here, the activity threshold for compounds from the public dataset was pChEMBL ≥ 6.5 to include all actives instead of only strong binders (pChEMBL > 8.5, which gave the best model performance). It was observed that the glycoside-like compounds cluster together in cluster 2 (**Figure 3.5**). Furthermore, the oxopyrrolidine-carboxamide compounds, which are also present in the public domain, are gathered in cluster 7. Cluster 4 mainly holds in-house compounds and includes the anti-histamine drug moxastine and antidepressant amitriptyline besides cloperastine. The differences in chemical structure, molecular weight, and ALogP of the clusters substantiate the possible existence of multiple binding sites. As a further example, cluster 6 differs considerably in ALogP from the other clusters. This suggests that the compounds in this cluster bind to a more hydrophilic site. The cluster centers and distribution of molecular weight, ALogP, number of hydrogen bond donors, and number of hydrogen bond acceptors for all clusters are shown in **Figure 3.6**. Additional pharmacological experiments, beyond the scope of this study, are warranted to further investigate the existence of multiple binding pockets in SGLT1. Attempts have been made to explore the binding sites of SGLT1 for substrates and inhibitor phloridzin.[35,36] Although the SGLT structure of vibrio parahaemolyticus has been used to generate hypotheses on SGLT1 binding pockets, the lack of an hSGLT1 structure hampers the detection of potential allosteric binding pockets.[37]

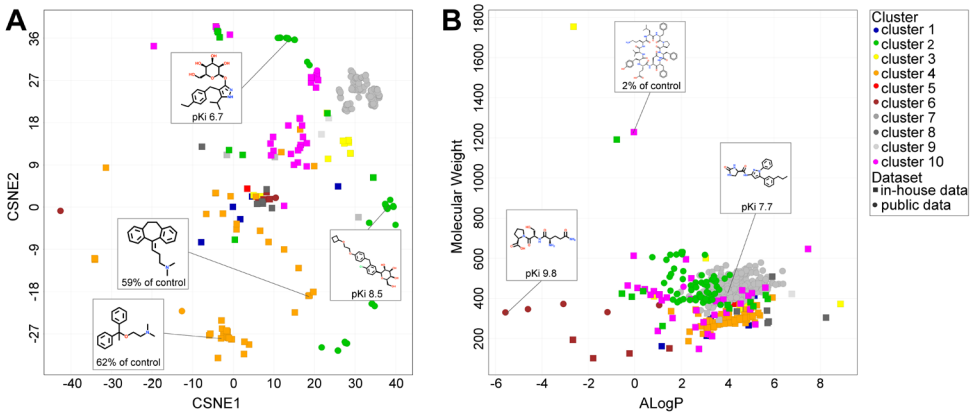


Figure 3.5. Clustering of hSGLT1 actives. Active hSGLT1 compounds in the training set clustered into ten chemical clusters (Tanimoto, FCFP6). Molecular structure and affinity (pKi for public data and % of (negative) control for in-house data) for representative cluster compounds are shown. In-house compounds with activity < 70% of (negative) control and public compounds with pChEMBL ≥ 6.5 were used in clustering. A) t-SNE plot of the chemical clusters. B) the molecular weight and ALogP distribution of compounds in the chemical clusters.

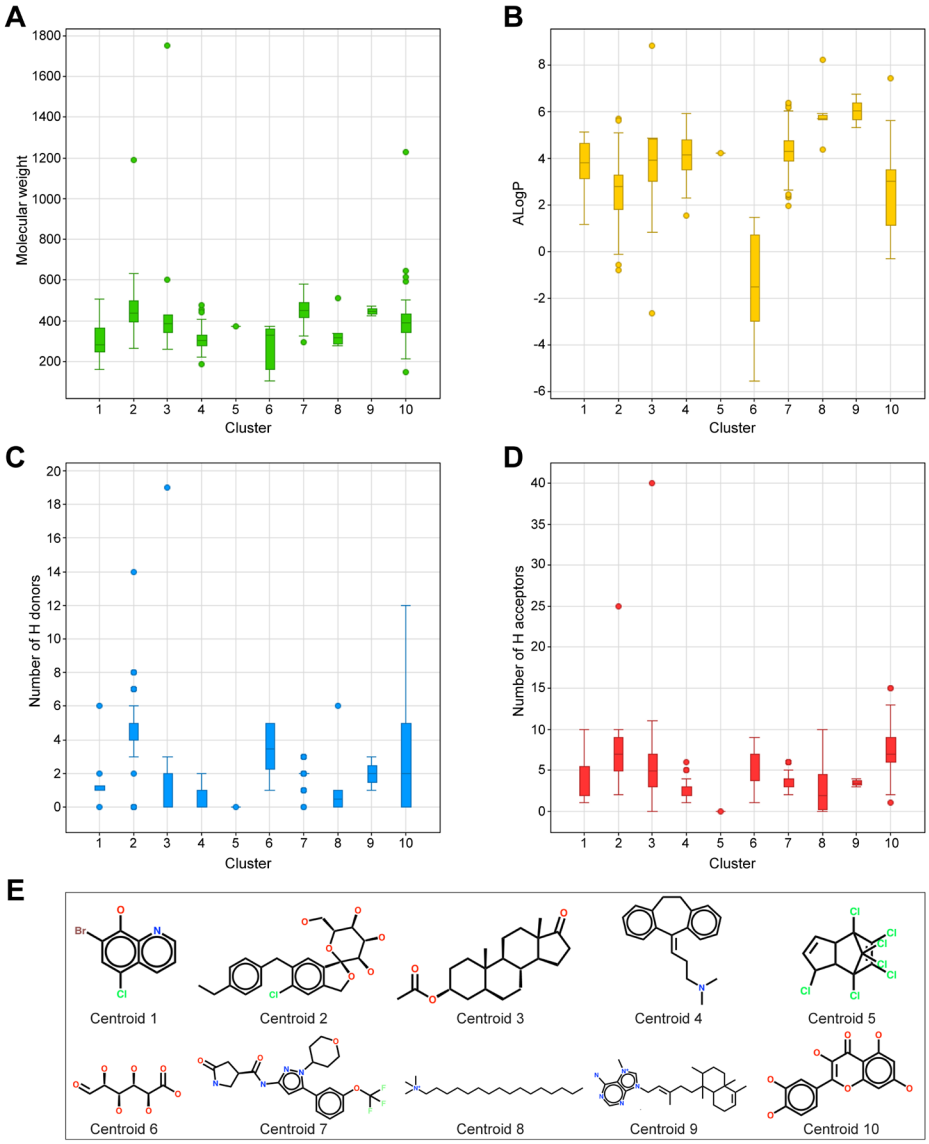


Figure 3.6. Cluster centers and distribution of physicochemical properties of hSGLT1 active compound clusters. Physicochemical properties of the ten clusters containing active hSGLT1 compounds. A) molecular weight. B) ALogP. C) number of hydrogen bond donors. D) number of hydrogen bond acceptors. E) molecular structures of the cluster centers.

Conclusion

PCM modeling appears a viable method to identify novel inhibitors for solute carrier hSGLT1 and likely any solute carrier protein. A predictive SGLT model was built with a MCC value of 0.49 ±0.05, estimated with 5-fold cross-validation. With the optimized model a hit rate of 38% was achieved when it was applied to screen for diverse molecules (Diverse set). In parallel, the model was used to boost identification of actives with a given chemotype (Cluster set). Although additional active compounds were identified, the data was too ambiguous to gain insight into the SAR of hSGLT1 inhibitors.

Diversity was found within the in-house dataset and differences were observed with the chemical space of the public dataset. Furthermore, the intrinsic variety in chemical structure of active compounds implies that there may be multiple binding sites at the transporter protein.

The novel identified inhibitors showed low similarity towards the training set and belong to the same chemical space of the in-house dataset, in contrast to the public dataset. Although the inhibitors were not optimized for specific drug delivery to the gastrointestinal tract, it is suggested that alterations (such as an increase in molecular weight and size) can make these inhibitors selective for intestinal hSGLT1.

Methods

Compounds and assay materials

DMEM-F12 (Biowest, Cat. No. L0092-500), DMEM (Lonza, BE12-604F/U1), Heat Inactivated Foetal Bovine Serum (HI-FBS, Biowest, Cat. No. S181H-500) and HBSS without Ca and Mg (HyClone, Cat. No. SH30588.01), DPBS (HyClone, Cat. No. SH30028.02), isopropanol (20842.312), clear-bottom black 96 well plates (Greiner, Cat. No. 655090) and polypropylene 96-well plates (Nunc, Cat. No. 151193) were all obtained from VWR (Amsterdam, the Netherlands). TrypLE Express (Gibco, Cat. No. 12605010), geneticin (Gibco, Cat. No. 10131027), D-glucose free DMEM (Gibco, Cat. No. 11966025), water soluble probenecid (Invitrogen, Cat. No. P36400), 5000 U/mL penicillin-streptomycin (Gibco, Cat. No. 15070063) were all ordered from Thermo Fisher Scientific (Breda, the Netherlands). 1-NBD-Glucose was custom synthesized by Mercachem (Nijmegen, the Netherlands). Bovine serum albumin (Cat. No. A8806), poly-L-lysine hydrobromide mol. wt. 30,000-70,000 (Cat. No. P2636), cell culture grade DMSO (Cat. No. D2650) were all acquired from Sigma-Aldrich Chemie (Zwijndrecht, the Netherlands). The hSGLT1 cDNA cloned in the pCMV6-neo vector was purchased from Origene Technologies (Rockville, USA, Cat. No. SC119918). The hSGLT2 cDNA was custom synthesized and cloned into the pcDNA3.1 vector by Thermo Fisher Scientific (Breda, the Netherlands). The experimentally tested Enamine screening compounds were acquired from Enamine (Kyiv, Ukraine).

Assay procedure

Two days in advance, CHO-hSGLT1 or CHO-hSGLT2 cells were seeded in maintenance medium (DMEM-F12 supplemented with 10% HI-FBS and 400 μg/mL geneticin) at 60.000 cells/well in clear-bottom black 96 well plates, pre-coated with 100 μg/mL poly-lysine. Cells were washed with 240 μL/well D-glucose free DMEM. Dilutions of test compounds and controls prepared

in D-glucose free DMEM with 350 μM 1-NBD-Glucose, 0.3% BSA, and 2 mM probenecid were added at 90 μL/well and placed in a humidified incubator at 37 °C with 5% CO_2 for 30 min. Subsequently cells were washed once with ice-cold DMEM-F12 and once with ice-cold HBSS, both at 240 μL/well. Finally, 1-NBD-Glucose was extracted from the cells with 100 μL/well isopropanol for 10 min at 600 rpm on an orbital shaker. Fluorescence was measured on a Flexstation 3 (Molecular Devices, San Jose, USA) with excitation at 445 nm, emission at 525 nm and cut off 515 nm. The uptake of 1-NBD-Glucose was normalized to the dynamic range between minimal inhibition (0.2% DMSO vehicle control) and maximal inhibition (100 μM phloridzin, >100x SGLT1/2 IC_{50}). Phloridzin is a strong inhibitor of SGLT1 and SGLT2 and was used as 0% reference, with 100% being no inhibition. A concentration of 100 μM phloridzin was used to ensure full SGLT1/2 inhibition. The Z-factor for the controls was determined and only data with Z > 0.4 (average Z SGLT1 assays: 0.8±0.1, average Z SGLT2 assays: 0.6±0.1) was used.[38]

Cytotoxicity assay

The cytotoxicity of compounds was tested with the ToxiLight bioassay kit (Lonza, obtained from VWR, Amsterdam, The Netherlands) according to the supplier's instructions. This non-destructive assay measures leakage of the enzyme AK from damaged cells into the CHO-hSGLT1/2 inhibition assay media, i.e. the degree of cytolysis. AK converts ADP into ATP and the enzyme luciferase subsequently catalyzes the formation of light from ATP and luciferin. Briefly, 20 mL of CHO-SGLT1/2 inhibition assay medium was added to 100 mL reconstituted AK detection reagent in white 96 wells Cellstar plates (Greiner bio-one, obtained from VWR, Amsterdam, The Netherlands) and incubated for 5 min at room temperature. Next, bioluminescence was measured on a FlexStation 3 Multi-Mode Microplate Reader (Molecular Devices, San Jose, USA) by 1 second integrated reading. Cytotoxicity was expressed as the percentage of bioluminescence of the 0.5% DMSO vehicle control which was set at 0%. The average cytotoxicity was calculated from biological replicates as indicated and average values > 20% were considered toxic (arbitrary threshold).

Dataset

Publicly available data from ChEMBL (version 23) was extracted for human SGLT1 (accession: P13866), human SGLT2 (P31639), and related proteins human SGLT3 (Q9NY91), rat SGLT1 (P53790), rat SGLT2 (P53792), mouse SGLT1 (Q9QXI6), mouse SGLT2 (Q923I7), and mouse SGLT3 (Q8R479). The retrieved compounds were standardized by removing salts, keeping the largest fragment, standardizing stereoisomers, standardizing charges, deprotonating bases, protonating acids, and optimizing the 2D structure by correcting bond lengths and angles. Activity values with confidence score 7 and 9 were kept and duplicate activity values were discarded based on activity standard unit ranking: $K_i > IC_{50} > EC_{50} > K_d$. For duplicate compounds with similar activity standard units (e.g. a compound with two Ki values), the average pChEMBL value was calculated.

An additional in-house dataset was provided by Unilever, Vlaardingen.[23] This dataset was based on the Spectrum Collection compound library (MicroSource Discovery Systems) extended with additional compounds that were similar to primary bioassay screening hits. This dataset consisted of compound activity data for hSGLT1 and hSGLT2. The activity was expressed as percentage

1-NBD-Glucose uptake compared to control at 50 μM, with control being the absence of inhibitor (=100%). Molecular structures were standardized in the same manner as the public data. The final dataset (public and in-house datasets combined, no duplicates) encompassed 3,686 unique compounds with 4,208 derived activities, of which 2,888 for hSGLT1.

Compound descriptors

Compounds were described using 512 FCFP6 fingerprint bits and the following physicochemical properties: molecular weight, ALogP, number of hydrogen bond acceptors, number of hydrogen bond donors, number of rotatable bonds, number of bridge bonds, and number of aromatic rings. Fingerprints and physicochemical descriptors were calculated in Pipeline Pilot (version 16.1.0).[39]

Protein descriptors

Protein sequences were aligned using whole sequence alignment in Clustal Omega (version 1.2.2).[40] Subsequently the sequences were converted to protein descriptors using Z-scales.[41] The first three Z-scales were implemented as protein descriptor as these were shown to perform well in previous work.[42] These three Z-scales include information on residue lipophilicity, size, and polarity.

Machine learning

Models were trained using the Random Forest R component in Pipeline Pilot (version 16.1.0). The number of trees was 500 and number of variables tried at each split was 38 (square root of the number of descriptors). Remaining settings were kept default.

T-Distributed Stochastic Neighbor Embedding

T-SNE was calculated on FCFP6 fingerprint descriptors that were converted to 2024 bits. The t-SNE component in Pipeline Pilot (version 18.1.0) was used to perform tSNE. The derived t-SNE values are represented by two components: CSNE1 and CSNE2.

Clustering of hSGLT1 actives to explore binding modes

hSGLT1 active compounds in the training set were clustered into ten clusters using the cluster molecules component in Pipeline Pilot (version 16.1.0). Compounds from the in-house set were included as 'active' when percentage of (negative) control was < 70%. Compounds from the public data set were termed 'active' when pChEMBL value ≥ 6.5.

Computational Hardware

Experiments were performed on a server running CentOS 6.9 equipped with a dual Xeon E-5 2630 v2 processor and 128 GB of RAM.

References

(1) Wood, I. S.; Trayhurn, P. Glucose Transporters (GLUT and SGLT): Expanded Families of Sugar Transport Proteins. 2003. https://doi.org/10.1079/BJN2002763.
(2) Tsujihara, K.; Hongu, M.; Saito, K.; Kawanishi, H.; Kuriyama, K.; Matsumoto, M.; Oku, A.; Ueta, K.; Tsuda, M.; Saito, A. Na+-Glucose Cotransporter (SGLT) Inhibitors as Antidiabetic Agents. 4. Synthesis and Pharmacological Properties of 4'-Dehydroxyphlorizin Derivatives Substituted on the B Ring. J. Med. Chem. 1999, 42, 5311–5324. https://doi.org/10.1021/jm990175n.
(3) Wright, E. M.; Loo, D. D.; Hirayama, B. A. Biology of Human Sodium Glucose Transporters. Physiol. Rev. 2011, 91, 733–794.
(4) Gorboulev, V.; Schürmann, A.; Vallon, V.; Kipp, H.; Jaschke, A.; Klessen, D.; Friedrich, A.; Scherneck, S.; Rieg, T.; Cunard, R.; et al. Na(+)-D-Glucose Cotransporter SGLT1 Is Pivotal for Intestinal Glucose Absorption and Glucose-Dependent Incretin Secretion. Diabetes 2012, 61, 187–196. https://doi.org/10.2337/db11-1029.
(5) Clar, C.; Gill, J. A.; Court, R.; Waugh, N. Systematic Review of SGLT2 Receptor Inhibitors in Dual or Triple Therapy in Type 2 Diabetes. BMJ Open 2012, 2, e001-007. https://doi.org/10.1136/bmjopen-2012-001007.
(6) Rosenstock, J.; Seman, L. J.; Jelaska, A.; Hantel, S.; Pinnetti, S.; Hach, T.; Woerle, H. J. Efficacy and Safety of Empagliflozin, a Sodium Glucose Cotransporter 2 (SGLT2) Inhibitor, as Add-on to Metformin in Type 2 Diabetes with Mild Hyperglycaemia. Diabetes, Obes. Metab. 2013, 15, 1154–1160. https://doi.org/10.1111/dom.12185.
(7) Komoroski, B.; Vachharajani, N.; Feng, Y.; Li, L.; Kornhauser, D.; Pfister, M. Dapagliflozin, a Novel, Selective SGLT2 Inhibitor, Improved Glycemic Control over 2 Weeks in Patients with Type 2 Diabetes Mellitus. Clin. Pharmacol. Ther. 2009, 85, 513–519. https://doi.org/10.1038/clpt.2008.250.
(8) Sands, A. T.; Zambrowicz, B. P.; Rosenstock, J.; Lapuerta, P.; Bode, B. W.; Garg, S. K.; Buse, J. B.; Banks, P.; Heptulla, R.; Rendell, M.; Cefalu, W. T.; Strumph, P. Sotagliflozin, a Dual SGLT1 and SGLT2 Inhibitor, as Adjunct Therapy to Insulin in Type 1 Diabetes. Diabetes Care 2015, 38, 1181–1188. https://doi.org/10.2337/dc14-2806.
(9) Rendell, M. S. Efficacy and Safety of Sotagliflozin in Treating Diabetes Type 1. Expert Opin. Pharmacother. 2018, 19, 307–315. https://doi.org/10.1080/14656566.2017.1414801.
(10) Masumoto, S.; Akimoto, Y.; Oike, H.; Kobori, M. Dietary Phloridzin Reduces Blood Glucose Levels and Reverses SGLT1 Expression in the Small Intestine in Streptozotocin-Induced Diabetic Mice. J. Agric. Food Chem. 2009, 57, 4651–4656. https://doi.org/10.1021/jf9008197.
(11) Rieg, T.; Vallon, V. Development of SGLT1 and SGLT2 Inhibitors. Diabetologia 2018, 61, 2079–2086. https://doi.org/10.1007/s00125-018-4654-7.
(12) Spatola, L.; Finazzi, S.; Angelini, C.; Dauriz, M.; Badalamenti, S. SGLT1 and SGLT1 Inhibitors: A Role to Be Assessed in the Current Clinical Practice. Diabetes Ther. 2018, 9, 427–430. https://doi.org/10.1007/s13300-017-0342-8.
(13) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. Nucleic Acids Res. 2017, 45, D945–D954. https://doi.org/10.1093/nar/gkw1074.
(14) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. Nucleic Acids Res. 2014, 42, D1083-90. https://doi.org/10.1093/nar/gkt1031.
(15) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100–D1107. https://doi.org/10.1093/nar/gkr777.
(16) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. Med. Chem. Commun. 2011, 2, 16–30. https://doi.org/10.1039/C0MD00165A.
(17) Lin, L.; Yee, S. W.; Kim, R. B.; Giacomini, K. M. SLC Transporters as Therapeutic Targets: Emerging Opportunities. Nat. Rev. Drug Discov. 2015, 14, 543–560. https://doi.org/10.1038/nrd4626.
(18) De Bruyn, T.; van Westen, G. J. P.; IJzerman, A. P.; Stieger, B.; de Witte, P.; Augustijns, P. F.; Annaert, P. P. Structure-Based Identification of OATP1B1/3 Inhibitors. Mol. Pharmacol. 2013, 83, 1257–1267.
(19) van Westen, G. J. P.; van den Hoven, O. O.; van der Pijl, R.; Mulder-Krieger, T.; de Vries, H.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. J. Med. Chem. 2012, 55, 7010–7020. https://doi.org/10.1021/jm3003069.
(20) Tresadern, G.; Trabanco, A. A.; Pérez-Benito, L.; Overington, J. P.; van Vlijmen, H. W. T.; van Westen, G. J. P. Identification of Allosteric Modulators of Metabotropic Glutamate 7 Receptor Using Proteochemometric Modeling. J. Chem. Inf. Model. 2017, 57, 2976–2985. https://doi.org/10.1021/acs.jcim.7b00338.

(21) van Westen, G. J. P.; Wegner, J. K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. PLoS One 2011, 6, e27518. https://doi.org/10.1371/journal.pone.0027518.
(22) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicabilty Domain Estimation by Projection of the Training Set Descriptor Space: A Review. Altern. Lab. Anim. 2005, 33, 445–459.
(23) Oranje, P.; Gouka, R.; Burggraaff, L.; Vermeer, M.; Chalet, C.; Duchateau, G.; van der Pijl, P.; Geldof, M.; de Roo, N.; Clauwaert, F.; Vanpaeschen, T.; Nicolaï, J.; de Bruyn, T.; Annaert, P.; IJzerman, A. P.; van Westen, G. J. P. Novel Natural and Synthetic Inhibitors of Solute Carriers SGLT1 and SGLT2. Pharmacol. Res. Perspect. 2019, 7, e00504. https://doi.org/10.1002/prp2.504.
(24) MicroSource Discovery Systems. Spectrum Collection http://www.msdiscovery.com/spectrum.html (accessed Nov 25, 2015).
(25) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. J. Mach. Learn. Res. 2008, 9, 2579–2605.
(26) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 2001, 46, 3–26.
(27) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. J. Comput. Aided. Mol. Des. 2015, 29, 885–896. https://doi.org/10.1007/s10822-015-9860-5.
(28) Enamine. Enamine HTS Collection https://enamine.net (accessed Oct 1, 2017).
(29) Charmot, D. Non-Systemic Drugs: A Critical Review. Curr. Pharm. Des. 2012, 18, 1434–1445. https://doi.org/10.2174/138161212799504858.
(30) Secker, P. F.; Beneke, S.; Schlichenmaier, N.; Delp, J.; Gutbier, S.; Leist, M.; Dietrich, D. R. Canagliflozin Mediated Dual Inhibition of Mitochondrial Glutamate Dehydrogenase and Complex I: An off-Target Adverse Effect. Cell Death Dis. 2018, 9, 226. https://doi.org/10.1038/s41419-018-0273-y.
(31) Goodwin, N. C.; Ding, Z.-M.; Harrison, B. A.; Strobel, E. D.; Harris, A. L.; Smith, M.; Thompson, A. Y.; Xiong, W.; Mseeh, F.; Bruce, D. J.; et al. Discovery of LX2761, a Sodium-Dependent Glucose Cotransporter 1 (SGLT1) Inhibitor Restricted to the Intestinal Lumen, for the Treatment of Diabetes. J. Med. Chem. 2017, 60, 710–721. https://doi.org/10.1021/acs.jmedchem.6b01541.
(32) Powell, D. R.; Smith, M. G.; Doree, D. D.; Harris, A. L.; Greer, J.; DaCosta, C. M.; Thompson, A.; Jeter-Jones, S.; Xiong, W.; Carson, K. G.; et al. LX2761, a Sodium/Glucose Cotransporter 1 Inhibitor Restricted to the Intestine, Improves Glycemic Control in Mice. J. Pharmacol. Exp. Ther. 2017, 362, 85–97.
(33) Lehmann, A.; Hornby, P. J. Intestinal SGLT1 in Metabolic Health and Disease. Am. J. Physiol. Liver Physiol. 2016, 310, G887–G898. https://doi.org/10.1152/ajpgi.00068.2016.
(34) Poulsen, S. B.; Fenton, R. A.; Rieg, T. Sodium-Glucose Cotransport. Curr. Opin. Nephrol. Hypertens. 2015, 24, 463–469. https://doi.org/10.1097/MNH.0000000000000152.
(35) Lostao, M. P.; Hirayama, B. A.; Loo, D. D. F.; Wright, E. M. Phenylglucosides and the Na+/Glucose Cotransporter (SGLT1): Analysis of Interactions. J. Membr. Biol. 1994, 142, 161–170. https://doi.org/10.1007/BF00234938.
(36) Bisignano, P.; Ghezzi, C.; Jo, H.; Polizzi, N. F.; Althoff, T.; Kalyanaraman, C.; Friemann, R.; Jacobson, M. P.; Wright, E. M.; Grabe, M. Inhibitor Binding Mode and Allosteric Regulation of Na+-Glucose Symporters. Nat. Commun. 2018, 9, 5245. https://doi.org/10.1038/s41467-018-07700-1.
(37) Watanabe, A.; Choe, S.; Chaptal, V.; Rosenberg, J. M.; Wright, E. M.; Grabe, M.; Abramson, J. The Mechanism of Sodium and Substrate Release from the Binding Pocket of VSGLT. Nature 2010, 468, 988–991.
(38) Zhang, J.-H.; Chung, T. D. Y.; Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J. Biomol. Screen. 1999, 4, 67–73. https://doi.org/10.1177/108705719900400206.
(39) Dassault Systèmes BIOVIA. BIOVIA Pipeline Pilot. Dassault Systèmes: San Diego 2016.
(40) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. Mol. Syst. Biol. 2011, 7, 539.
(41) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J. Med. Chem. 1998, 41, 2481–2491. https://doi.org/10.1021/jm9700575.
(42) van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. J. Cheminform. 2013, 5, 42. https://doi.org/10.1186/1758-2946-5-42.

.

Supporting Information

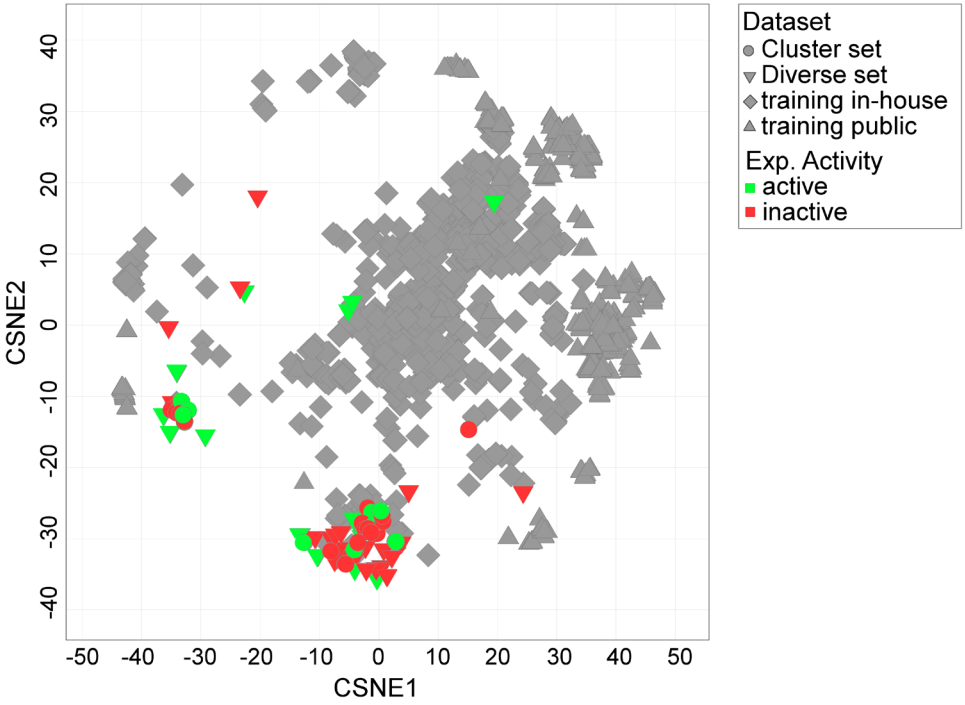


Figure S3.1. T-SNE representation of the chemical space of actives and inactives of selected compounds compared to the training set. Experimental activities of compounds in the Diverse set and Cluster set compared to the training set (grey).

Table S3.2. Bioactivities, cytotoxicities and Tanimoto similarities of the Diverse set.

Structure	Compound ID	SGLT1 activity (% of control)		Activity class	SGLT2 activity (% of control)		Similarity towards training (Tanimoto)	Toxicity (AK secretion)				
		Average	St.dev		Average	St.dev		Average	St.dev	n	n1	n2
CN(CCOc1c(C)cccc1C)Cc2ccccc2N3CCOCC3	Z163972344	24	1.1	active	29	5.5	0.271	-8.80	2.10	2	-10.3	-7.3
N\C(=N\OCc1ccccc1)\c2ccccc2CN3CCCCC3	Z915954934	28	3.7	active	42	1.7	0.431	2.24	10.38	2	9.6	-5.1
CN(C)CCCN1\C(=N/c2ccccc2)\SC=C1c3ccccc3	Z56792856	35	4.1	active	29	3.6	0.328	7.10	23.30	2	23.6	-9.4
CC(N(C)S(=O)(=O)N1CCCCCC1)c2cccc(c2)[N+](=O)[O-]	Z237884404	39	3.0	active	35	5.0	0.295	-14.70		1		
CN(C)CCN(Cc1occc1)c2ccc(Cl)c(n2)C#N	Z1416510792	43	9.2	active	45	15.6	0.614	25.30		1		
CC1CCCN(CCC(C#N)(c2ccccc2)c3ccccc3)C1	Z137703320	47	6.9	active	28	10.9	0.321	4.71	15.42	5		
Fc1ccc(CN(CCc2ccccn2)CC#Cc3ccccc3)cc1	Z1844922248	49	7.0	active	48	12.1	0.366	-9.20		1		
OC1(CC2CCC(C1)N2CCOC3CCCCO3)c4ccccc4	Z2677914007	51	3.1	active	50	6.1	0.329	8.10	14.50	2	18.4	-2.1
O=C(CN1CCC2CCCCC2C1)Nc3cccc(c3)S(=O)(=O)N4CCCCC4	Z46498483	56	3.8	active	27	5.8	0.269	8.70	18.60	2	21.9	-4.5
CCOC(=O)C1CCCCN1S(=O)(=O)c2ccc(cc2)C(C)(C)C	Z45484979	57	0.7	active	42	2.9	0.321	-0.02	9.05	5		
CC1=CSC(=O)N1CCC(=O)N2CCN(Cc3ccccc3)C4(CCCCC4)C2	Z1268385265	61	12.7	active	69	15.0	0.265	-3.49	5.79	2	0.6	-7.6
Fc1ccc(cc1Cl)C2CN(CCS(=O)(=O)c3ccccc3)CCO2	Z1070584838	62	6.1	active	61	4.7	0.343	3.59	10.50	4		
CN(Cc1ccccc1NC(=O)N2CCCC(C2)N3CCCC3)C4CCCCC4	Z594462480	63	5.6	active	54	5.2	0.312	2.39	10.75	4		
CN(CC(=O)N1CC(c2ccccc2)c3ccccc3C1)S(=O)(=O)c4ccc(cc4)C#N	Z289555820	63	18.6	active	60	2.1	0.257	-3.35	12.25	2	-12.0	5.3
Cc1c(C)n(C2CCCCC2)c(NC(=O)CN3CCCC3C(C#N)c4ccccc4)c1C#N	Z1301318288	65	8.6	active	56	3.2	0.261	10.55	2.19	2	12.1	9.0
CN(CCOC(c1ccccc1)c2ccccc2)S(=O)(=O)C	Z1642825375	72	4.4	inactive	67	15.7	0.743	14.10	14.70	2	24.5	3.7
O=S(=O)(N1CCC1c2ccccc2)c3cccc4N=S=Nc34	Z741018970	73	6.6	inactive	86	4.7	0.232	7.00	5.60	4		

Structure	Compound ID	SGLT1 activity (% of control)		Activity class	SGLT2 activity (% of control)		Similarity towards training (Tanimoto)	Toxicity (AK secretion)				
		Average	St.dev		Average	St.dev		Average	St.dev	n	n1	n2
CCS(=O)(=O)c1ccccc1C(=O)N2CCCN(CC3CCCCC3)CC2	Z783393074	76	13.0	inactive	77	6.6	0.297	10.80	12.12	5		
CN(CCC#N)Cc1c(C)nn(Cc2ccccc2Cl)c1C	Z103956902	76	4.6	inactive	74	7.1	0.258	-6.47	15.72	5		
CC1CC(CCN1Cc2ccccc2)N(C)Cc3cnns3	Z2161555222	76	3.0	inactive	80	4.7	0.316	-1.30	9.99	5		
CCC1CN2CCCC2CN1Cc3ccc(OC(F)F)cc3	Z1318121286	78	7.6	inactive	69	10.2	0.348	7.14	11.83	5		
CN(C1CCOC1)S(=O)(=O)c2ccccc2C(=O)c3ccccc3	Z1725346307	82	2.7	inactive	80	13.4	0.295	7.86	18.23	4		
CCC(CN1CCC(CC1)OCC2CCOC2)c3ccccc3	Z1661722958	83	5.2	inactive	78	13.2	0.322	0.30	12.95	5		
N#Cc1ccccc1c2ccc(CN3CCC4(CC3)OCCO4)cc2	Z368852630	83	5.4	inactive	67	15.4	0.345	-4.79	4.50	5		
CCN(CC)CCN1CN(Cc2ccccc2)c3nc4ccccc4n3C1	Z104342888	86	5.6	inactive	66	1.7	0.263	-3.10	12.27	5		
CCCCN(CC#Cc1ccc(cc1)C(F)(F)F)C2CCS(=O)(=O)C2	Z1332354136	86	7.6	inactive	67	7.3	0.286	-9.27	9.75	5		
CCCCC(N1CCCS1(=O)=O)C(=O)Nc2cc(C)c(F)cc2Cl	Z2151290710	88	2.9	inactive	93	8.9	0.397	2.00	13.77	4		
C(C1CCCN(Cc2cccc(c2)n3cccn3)C1)N4CCOCC4	Z1568339019	89	10.2	inactive	80	4.8	0.279	6.21	12.79	5		
FC(F)(F)C(N1CCN(CC2OCCO2)CC1)c3ccccc3	Z1612085883	89	1.9	inactive	84	11.7	0.259	7.06	21.91	4		
C(C1CCOCC1)N(Cc2ccccc2)Cc3cn4cccnc4n3	Z1176622491	90	7.3	inactive	93	10.1	0.250	-0.12	9.90	4		
N#CC1CCC(C1)N2CCC(CN3CCc4ccccc34)CC2	Z1912611670	90	11.4	inactive	64	10.5	0.286	-9.32	18.88	5		
C\C(=C\c1ccccc1)\C=N\N2CCN(Cc3ccccc3Cl)CC2	Z56796910	91	4.5	inactive	59	12.8	0.276	8.70		1		
N#Cc1c(CN2CCC(=CC2)c3ccccc3)cn4ccccc14	Z220381090	94	14.7	inactive	84	16.2	0.250	-0.96	10.32	5		
CC1(C)CC(N(Cc2ccccc2)C(=O)CCCS(=O)(=O)N3CCCC3)C(C)(C)O1	Z1609106419	95	1.0	inactive	89	12.8	0.271	8.73	10.20	4		
Cn1cc(CN2CCCN(CC2)C3Cc4ccccc4C3)cn1	Z1176394041	96	6.3	inactive	79	10.2	0.254	4.62	10.59	4		
CN1CCCC1CCSc2nnc(C)n2Cc3ccccc3	Z1526821488	97	10.5	inactive	82	3.8	0.338	9.66	21.34	4		

Structure	Compound ID	SGLT1 activity (% of control)		Activity class	SGLT2 activity (% of control)		Similarity towards training (Tanimoto)	Toxicity (AK secretion)				
		Average	St.dev		Average	St.dev		Average	St.dev	n	n1	n2
C(\C=C\c1ccccc1)N2CCN(CC2)C(c3ccccc3)c4ccc(cc4)c5ccccc5	Z1172208353	103	7.2	inactive	84	10.9	0.667	12.26	18.33	5		
O=C(CN1CCN(CC1)C(=O)\C(=C\c2ccccc2)\c3ccccc3)N4CCOCC4	Z73457041	114	5.6	inactive	88	15.2	0.273	-0.63	20.90	5		
S=C1N=C(NN1CN2CCCCC2)\C=C\c3ccccc3	Z238423340	115	10.5	inactive	107	9.5	0.288	-33.53	10.44	5		
Clc1nc2sccn2c1\C=N\N3CCN(CC3)C(c4ccccc4)c5ccccc5	Z49982582	122	6.6	inactive	109	13.6	0.286	15.01	18.31	5		

Table S3.3. Bioactivities, cytotoxicities and Tanimoto similarities of the Cluster set.

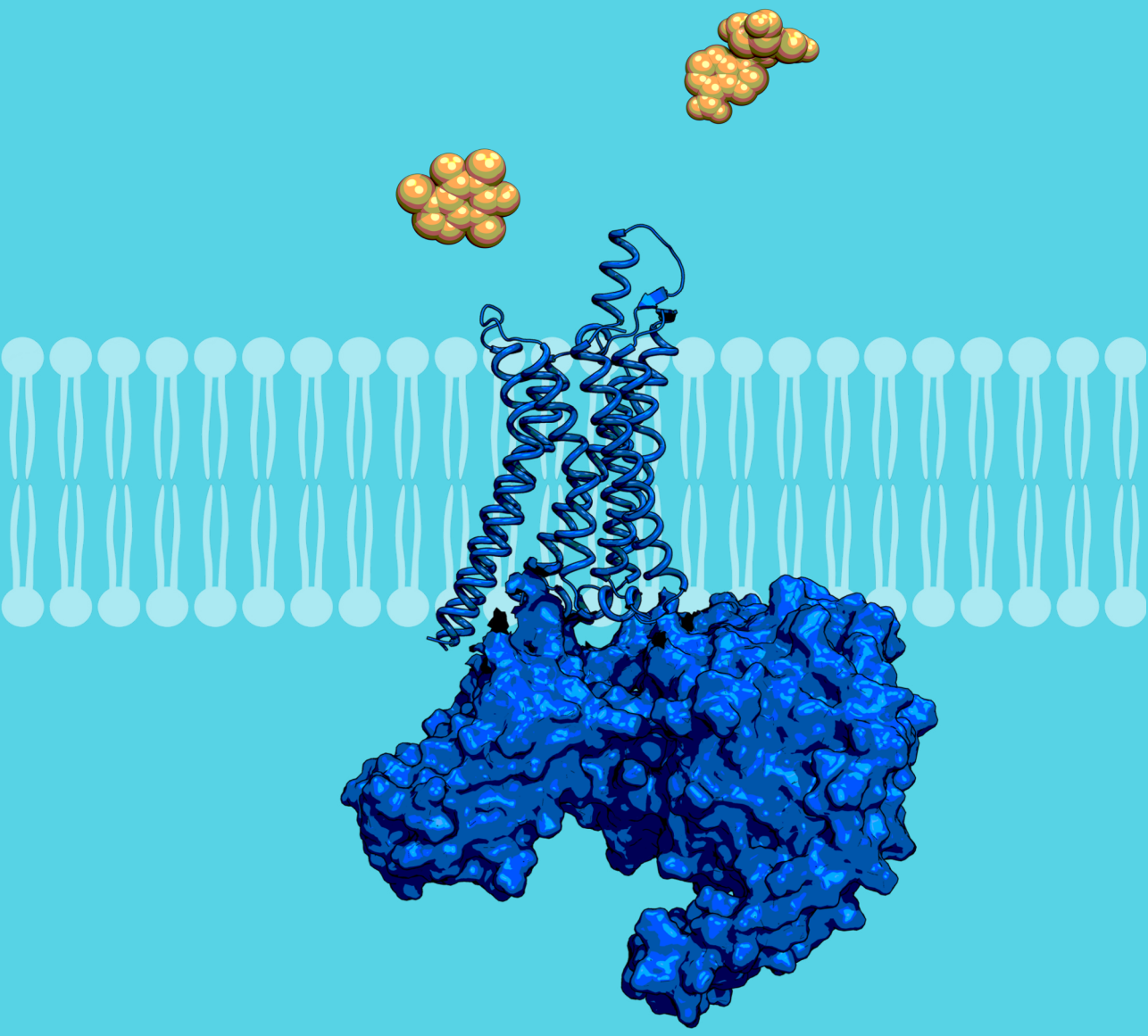
Structure	Compound ID	Reference compound	SGLT1 activity (% of control); n=2				Activity class	SGLT2 activity (% of control), n=2				Similarity towards training (Tanimoto)	Toxicity (AK secretion)		
			Average	St.dev	n1	n2		Average	St.dev	n1	n2		Average	St.dev	n
Clc1ccc(cc1)C(N2CCN(CCOCC(=O)N3CCCCC3)CC2)c4ccccc4	Z827380946	Cloperastine	31	11.0	22.8	38.9	active	30	5.5	25.9	33.6	0.370	-0.27	8.13	3
Cc1ccc(cc1)C(OCC(O)CN2CCCCC2)c3ccccc3	Z31367782	Cloperastine	36	4.4	33.3	39.6	active	35	0.5	34.4	35.1	0.492	-10.06	14.06	4
CC1CCCN(CC(O)COC(c2ccccc2)c3ccccc3)C1	Z31371621	Cloperastine	37	3.3	34.8	39.4	active	27	2.9	29.3	25.3	0.419	-5.72	8.22	3
CC(C)N(CC#CCC(O)(C1CCCC1)c2ccccc2)Cc3ccccc3	Z56906862	Trihexyphenidyl	38	5.4	41.6	34.0	active	25	5.0	21.5	28.5	0.441	-10.40		1
CC(CN1CCCCC1)OC(=O)C(c2ccccc2)c3ccccc3	Z204157928	Trihexyphenidyl	41	7.3	36.3	46.6	active	27	4.1	24.2	30.1	0.322	-18.49	10.29	4
CC(C)c1cccc(C(C)C)c1NC(=O)CN2CCCCCC2	Z46160496	Bupivacaine	41	4.3	43.7	37.7	active	55	2.1	56.6	53.6	0.414	-8.60		1
Cc1ccccc1C(OCC(O)CN2CCCCC2)c3ccccc3	Z31367784	Cloperastine	43	7.3	37.4	47.8	active	38	0.6	37.6	38.4	0.459	-14.87	17.28	3
CN(CC(O)COC(c1ccccc1)c2ccccc2)C3CCCCC3	Z31370217	Cloperastine	45	10.0	38.0	52.9	active	24	7.0	19.2	29.0	0.365	15.30	16.73	4
CC1CCCCN1CC(=O)Nc2ccccc2C(=O)c3ccccc3	Z46224544	Bupivacaine	45	9.6	38.2	51.8	active	20	0.1	20.1	20.3	0.403	-1.30		1
CC(C)CCSCC(=O)N1CC2CCCN2Cc3ccccc13	Z817504494	Bepridil	45	3.3	47.4	42.7	active	27	11.9	35.0	18.1	0.305			0
CN(CCC1CCCCC1)CC(=O)Nc2c(C)cccc2C	Z1003034804	Bupivacaine	55	1.0	54.5	55.9	active	41	1.3	41.8	40.1	0.426	1.67	8.02	4
CC(N1CCCCCCC1)C(=O)Nc2ccccc2C(F)(F)F	Z124096378	Bupivacaine	57	6.9	61.7	52.0	active	45	2.7	46.9	43.0	0.413	0.78	4.98	4

Structure	Compound ID	Reference compound	SGLT1 activity (% of control); n=2				Activity class	SGLT2 activity (% of control), n=2				Similarity towards training (Tanimoto)	Toxicity (AK secretion)		
			Average	St.dev	n1	n2		Average	St.dev	n1	n2		Average	St.dev	n
CC1CN(CCOC(c2ccccc2)c3ccccc3)CCC1=O	Z1751527676	Cloperastine	62	8.6	55.5	67.6	active	56	10.6	48.4	63.4	0.518	-15.40		1
CCOC(=O)C1CCCCN1CC(=O)N2CCCc3ccccc23	Z105569118	Bepridil	64	4.0	60.8	66.4	active	76	4.5	79.4	73.0	0.286	15.96	30.89	3
O=C(CN1CCCCCC1)N(Cc2ccccc2)c3ccccc3	Z199542950	Bepridil	65	0.4	65.7	65.1	active	53	4.2	49.6	55.6	0.394	-0.80		1
COCC(C)N1CCN(CCCOCc2ccccc2)CC1	Z1665888845	Bepridil	71	6.8	66.2	75.7	inactive	74	1.6	74.8	72.6	0.324	5.75	10.97	3
NC(=O)C1CCCCN1CC(=O)N(Cc2ccccc2)Cc3ccccc3	Z46442988	Bepridil	71	18.0	83.4	57.8	inactive	78	14.4	88.6	68.2	0.284	4.60		1
C(CN1CCC(CC1)OCc2ccccc2)Oc3ccccc3	Z196036526	Bepridil	72	1.9	70.7	73.4	inactive	47	0.7	46.1	47.1	0.301	4.82	14.47	3
CC(N1CCCCC1)C(=O)Nc2cc(C)ccc2C(=O)c3ccccc3	Z2217101732	Bupivacaine	74	8.1	67.8	79.3	inactive	43	7.2	47.9	37.7	0.403	-4.30		1
CC1CCCCN1CC(=O)Nc2ccccc2C(F)(F)F	Z46224455	Bupivacaine	75	2.4	77.1	73.8	inactive	47	11.5	55.3	39.0	0.422	-2.08	6.50	4
Cc1ccccc1NC(=O)CCN2CCCCC2C3OCCO3	Z1298879676	Bupivacaine	77	14.0	87.0	67.4	inactive	74	18.0	86.8	61.3	0.426	7.10		1
CCN(C)CC1CCN(C1)C(=O)CSC(c2ccccc2)c3ccc(Cl)cc3	Z1276034798	Cloperastine	78	3.1	80.0	75.7	inactive	55	5.9	59.4	51.0	0.435			0
CC(CCN1CC(C)C(C1)N2CCOCC2)c3ccccc3	Z1897146788	Bepridil	82	2.2	83.9	80.8	inactive	77	7.9	71.8	83.0	0.280	-6.70		1
CC1CCCCN1CC2CCCN(C2)C(=O)Nc3c(C)cccc3C	Z1230260964	Bupivacaine	83	1.7	81.3	83.7	inactive	82	12.7	91.3	73.3	0.449			0
Fc1ccccc1CCN(CCC#N)CCN2CCOCC2	Z1704430244	Cloperastine	84	8.4	90.0	78.1	inactive	87	10.3	94.4	79.8	0.343	0.50		1

Structure	Compound ID	Reference compound	SGLT1 activity (% of control); n=2				Activity class	SGLT2 activity (% of control), n=2				Similarity towards training (Tanimoto)	Toxicity (AK secretion)		
			Average	St.dev	n1	n2		Average	St.dev	n1	n2		Average	St.dev	n
CCCCC(N1CCC(CC1)OCc2ccccc2)C(=O)N	Z1084070528	Bepridil	85	18.0	97.8	72.3	inactive	89	3.6	91.4	86.3	0.284			0
Cc1cccc(C)c1NC(=O)CCCN2CCN(CC#C)CC2	Z1535223420	Bupivacaine	85	13.0	93.7	75.8	inactive	92	0.0	92.2	92.2	0.433			0
C[Si](C)(CN1CCCC(CCC(=O)N)C1)c2ccccc2	Z1912600813	Trihexyphenidyl	86	3.3	88.8	84.1	inactive	79	9.5	85.4	72.0	0.328	8.57	6.24	4
C(CN1CCC(CN2CCCCCC2)CC1)Cc3ccccc3	Z436106106	Trihexyphenidyl	88	0.8	87.5	88.6	inactive	82	8.3	75.7	87.4	0.357	-11.90		1
CCN(Cc1ccccc1)c2ccc(\C=N\N3CCN(C)CC3)cc2	Z49564244	Bepridil	91	3.0	88.6	92.9	inactive	93	3.9	89.8	95.3	0.361	-22.37	24.25	4
Fc1ccc(cc1)C(N2CCC(CN3CCOCC3)CC2)c4ccccc4	Z1318177320	Cloperastine	93	20.0	106.7	79.1	inactive	54	0.1	54.6	54.4	0.422			0
O=C(N1CCN(CCOC(c2ccccc2)c3ccccc3)CC1)c4ccccc4	Z195623196	Cloperastine	95	9.7	101.6	87.9	inactive	72	1.9	73.0	70.4	0.474	-14.70		1
NC(=O)C(N1CCC(CN2CCCC2)CC1)c3ccccc3	Z1103127942	Trihexyphenidyl	98	14.0	87.9	107.7	inactive	89	10.0	82.2	96.3	0.322	-13.70		1
NC(C(=O)N1CCCC(CN2CCCCCC2)C1)c3ccccc3	Z1459737806	Trihexyphenidyl	98	16.0	86.9	109.3	inactive	101	7.5	106.7	96.1	0.333	-7.90		1
OC(CN1CCC(CN2CCCCCC2)CC1)c3ccccc3Cl	Z416806614	Trihexyphenidyl	98	5.6	93.8	101.6	inactive	79	1.8	78.1	80.6	0.317	-11.30		1
CC1CN2CCCCC2CN1CC(=O)Nc3ccc(C)cc3	Z1223130644	Bupivacaine	104	2.3	102.3	105.5	inactive	95	15.1	84.7	106.1	0.415	-5.70		1
CCOCC1CCN(C1)C(=O)C(CN2CCCC2)c3ccccc3	Z2229127630	Bepridil	112	6.6	107.2	116.5	inactive	92	3.9	89.0	94.5	0.278	8.28	13.99	4

Chapter four

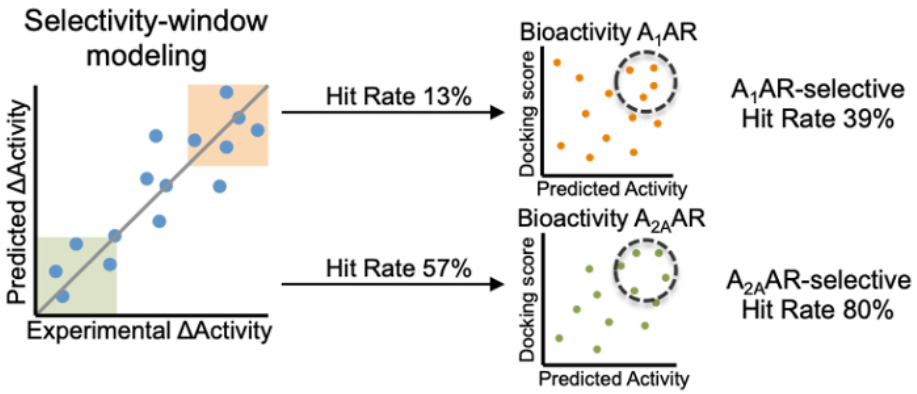
Quantitative Prediction of Selectivity between the A_1 and A_{2A} Adenosine Receptors



Lindsey Burggraaff, Herman W.T. van Vlijmen, Adriaan P. IJzerman, and Gerard J.P. van Westen.
Journal of Cheminformatics. 2020. doi: 10.1186/s13321-020-00438-3.

Abstract

The development of drugs is often hampered due to off-target interactions leading to adverse effects. Therefore, computational methods to assess the selectivity of ligands are of high interest. Currently, selectivity is often deduced from bioactivity predictions of a ligand for multiple targets (individual machine learning models). Here we show that modeling selectivity directly, by using the affinity difference between two drug targets as output value, leads to more accurate selectivity predictions. We test multiple approaches on a dataset consisting of ligands for the A_1 and A_{2A} adenosine receptors (among others classification, regression, and we define different selectivity classes). Finally, we present a regression model that predicts selectivity between these two drug targets by directly training on the difference in bioactivity, modeling the selectivity-window. The quality of this model was good as shown by the performances for 5-fold cross-validation: ROC A_1AR-selective 0.88 ±0.04 and ROC A_{2A}AR-selective 0.80 ±0.07. To increase the accuracy of this selectivity model even further, inactive compounds were identified and removed prior to selectivity prediction by a combination of statistical models and structure-based docking. As a result, selectivity between the A_1 and A_{2A} adenosine receptors was predicted effectively using the selectivity-window model. The approach presented here can be readily applied to other selectivity cases.



Introduction

Computational modeling of small molecules in drug discovery is typically focused on modeling their binding affinity or bioactivity. These models can be used to identify active compounds in silico, or to rationalize which chemical groups are correlated with bioactivity. Quantitative structure-activity relationship (QSAR) models can be applied to model compound activity for single proteins, whereas proteochemometrics (PCM) can be applied to model activity for multiple proteins in one single model.[1,2] Next to statistical models (e.g. machine learning), structure-based models are used to predict and rationalize compound activity. Methods such as docking, molecular dynamics, and free-energy perturbation (FEP) are widely applied to study binding and bioactivity.[3-5]

Although modeling of activity is essential in drug discovery, these models often do not take target selectivity into account. The ability to control promiscuity of potential drug candidates is crucial as off-target effects can be avoided this way. Whereas selective drugs are designed to be non-promiscuous, polypharmacological drugs are designed to interact with multiple targets.[6] The development of both polypharmacological and selective drugs requires predictions for more than one target. Polypharmacology and selectivity can both be modeled by machine learning or structure-based approaches that predict the affinity of compounds on multiple targets separately. The resulting bioactivity profile can subsequently be applied to match the desired on-target(s) and to avoid off-target effects. However, this indirect way of predicting selectivity based on predicted bioactivities requires multiple model predictions to calculate one feature: the activity difference of a compound for one target over the other.

Here we explore selectivity modeling for the adenosine receptors, which are members of the class A G protein-coupled receptors (GPCRs). The adenosine receptor family, existing of subtypes A_1, A_{2A}, A_{2B}, and A_3, is involved in many physiological processes including cardiac control and inflammation.[7] The A_1 and A_{2A} adenosine receptors (A_1AR and A_{2A}AR) both control cyclic adenosine-5'-monophosphate (cAMP) levels in the cell. Activation of A_1AR results in decreased cAMP levels, whereas A_{2A}AR activation increases cAMP levels.[8] These contrary effects justify a need for computational models that can predict selectivity between these two subtypes. A novel method to model selectivity is presented in this study: namely to train machine learning models directly on the differences between experimentally determined affinities.

For the A_{2A}AR many crystal structures are available in the Protein Data Bank.[9] More recently, protein crystal and cryo-EM structures for the A_1AR have been obtained also.[10-12] The availability of structures for both proteins allows for a structure-based comparison of the subtypes to gain insights into selectivity. Previous studies revealed differences between the protein structures that correspond with ligand selectivity of specific chemical structures.[10,11,13] For example, the A_{2A}AR-selectivity of reference antagonist ZM241385 could be explained by a combination of three structural differences between the A_1AR and A_{2A}AR: a salt bridge at the binding pocket entry, a hydrophobic pocket in the A_1AR, and a (de)stabilized water network.[13] Furthermore, the A_1AR-selectivity of xanthine-based antagonists with a bulky substituent has shown to be caused by steric hindrance in the A_{2A}AR by residue Met270$^{7.35}$ (Thr270$^{7.35}$ in A_1AR).[11]

In addition to the availability of crystal structures, many small molecule ligands have been experimentally tested for their activity on one or multiple adenosine receptors. This data has

already been exploited to train bioactivity models using classification or continuous statistical models.[14,15] However, direct modeling of selectivity in the adenosine receptors using statistical models has not yet been reported. Statistical models to predict selectivity are faster than predicting selectivity using time-consuming FEP approaches.[5] However, structure-based methods can give additional mechanistic information on binding and selectivity of compounds, and are in principle not dependent on available bioactivity data of ligands.

In this research a combined approach of QSAR modeling and structure-based docking is presented to model bioactivity for the A_1AR and $A_{2A}AR$. Moreover, the selectivity between the A_1AR and $A_{2A}AR$ is predicted directly by training on affinity differences (selectivity-window). This is in contrast to methods that were reported up till now that deduced selectivity from predicted bioactivities of separate models. Furthermore, by training a continuous model (regression), the degree of selectivity was calculated in addition to a selectivity class with predefined thresholds. Finally, to enhance the performance of the selectivity models, statistical bioactivity models and structure-based docking were used to exclude inactive compounds.

Our study shows that continuous QSAR models can effectively predict selectivity between the A_1AR and $A_{2A}AR$. A model trained directly on the difference in affinity between the two proteins, the selectivity-window model, outperformed models that are generally used to predict selectivity: models that are trained on separate bioactivities for the A_1AR and $A_{2A}AR$.

Results

Datasets

Information on compound-protein interactions (e.g. binding affinity and efficacy) was collected for the human A_1AR and $A_{2A}AR$. Public bioactivity data was extracted from ChEMBL version 23[16] and supplemented with in-house data. As compounds with a ribose scaffold are often associated with agonistic activity and dicyanopyridines with partial agonism, these compounds were removed to generate an antagonist-focused dataset.[17,18] Bioactivity values were standardized to pActivity (conceptually similar to the pChEMBL value, an ensemble from $pK_i/IC_{50}/EC_{50}/K_d$ values[19]), while simultaneously combining data from different labs and assays. The data was subsequently used to compile the following compound datasets: an 'A_1AR bioactivity dataset' (containing bioactivities of compounds on the A_1AR), an '$A_{2A}AR$ bioactivity dataset' (containing bioactivities of compounds on the $A_{2A}AR$) and an '$A_1AR/A_{2A}AR$ dataset' (bioactivities of compounds tested on both the A_1AR and $A_{2A}AR$). The latter dataset included information on the selectivity of compounds. Compounds were termed 'selective' when the difference in activity between the two proteins was more than 100 fold (e.g., $A_{2A}AR$-selective when pActivity A_1AR = 6.5 and pActivity $A_{2A}AR \geq 8.5$). The $A_1AR/A_{2A}AR$ dataset consisted of five classes: non-binder (pActivity A_1AR and $A_{2A}AR < 6.5$), A_1AR-selective (pActivity $A_1AR \geq 6.5$ and selectivity $\geq$ 100 fold), $A_{2A}AR$-selective (pActivity $A_{2A}AR \geq 6.5$ and selectivity $\geq$ 100 fold), and dual binder (pActivity A_1AR and $A_{2A}AR \geq 6.5$ and selectivity $\leq$ 10 fold). Additionally, compounds that had measured bioactivities for both the A_1AR and $A_{2A}AR$, but did not fit into any of the classes of the $A_1AR/A_{2A}AR$ dataset, were termed 'semi-selective' compounds. Compounds that only had measured bioactivity for one receptor and were not present in the $A_1AR/A_{2A}AR$ dataset, but were included in either the A_1AR bioactivity dataset or $A_{2A}AR$ bioactivity dataset were termed 'single points'. The distribution of activities in the different datasets was comparable (**Table 4.1**)

Table 4.1. Dataset characteristics: number of compounds, distribution of activities and chemical similarity within the dataset.

Dataset	Description	Total number of compounds	Activity (pActivity)				similarity (tanimoto FCFP4)
			protein	min	median	max	mean
A_1AR bioactivity dataset	Compounds with measured activity for the A_1AR	2774	A_1AR	4.05	6.43	10.52	0.18
$A_{2A}AR$ bioactivity dataset	Compounds with measured activity for the $A_{2A}AR$	3123	$A_{2A}AR$	4.00	6.91	11.00	0.18
$A_1AR/A_{2A}AR$ dataset	Compounds with measured activity for both the A_1AR and $A_{2A}AR$ with classification $A_1AR/A_{2A}AR$/dual/non-binder	1106	A_1AR	4.33	6.52	10.52	0.19
			$A_{2A}AR$	4.30	6.83	10.80	
Semi-selective compounds	Compounds with measured activity for both the A_1AR and $A_{2A}AR$ that do not fit into a class	855	A_1AR	4.37	6.37	10.02	0.20
			$A_{2A}AR$	4.34	7.09	10.38	

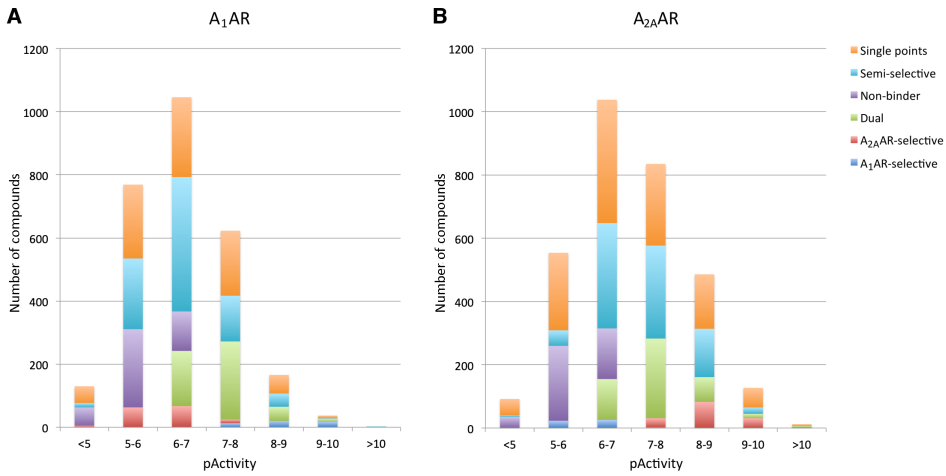


Figure 4.1. Distribution of activities of the different compound classes for the A_1AR (A) and $A_{2A}AR$ (B). The compounds from the A_1AR and $A_{2A}AR$ bioactivity datasets that did not belong to any of the classes of the $A_1AR/A_{2A}AR$ dataset are called "single points".

and normally distributed in both the A_1AR and $A_{2A}AR$ (**Figure 4.1**). It should be noted that the total number of A_1AR-selective compounds is about three times smaller than the number of $A_{2A}AR$-selective compounds (50 and 146 compounds, respectively).

Modeling $A_1AR/A_{2A}AR$ subtype selectivity using classification QSAR models

Several QSAR models were created to study selectivity. Firstly, subtype selectivity for the A_1AR and $A_{2A}AR$ was modeled using classification models. Additionally, non-selective compounds (dual binders) and non-binders were modeled. The following four models were constructed: a

2-class model (A_1AR-selective/A_{2A}AR-selective), two 3-class models (A_1AR-selective/A_{2A}AR-selective/dual inhibitors on one hand, and A_1AR-selective/A_{2A}AR-selective/non-binders on the other hand), and a 4-class model (A_1AR-selective/A_{2A}AR-selective/dual/non-binders). All models were validated with 5-fold cross-validation, using the same (chemically clustered) test sets per iteration for each model (for details see Methods section). The performance of the 2-class QSAR model was best for predicting A_1AR and A_{2A}AR selectivity (receiver operating characteristic (ROC) 0.87 ±0.06, and Matthews Correlation Coefficient (MCC) 0.40 ±0.13). Addition of dual and non-binder classes decreased model performance. ROC decreased to 0.76 ±0.06 and 0.64 ±0.09 for the A_1AR and to 0.62 ±0.09 and 0.65 ±0.09 for the A_{2A}AR. Likewise, MCC decreases to 0.22 ±0.15 and 0.00 ±0.07 for the A_1AR and to 0.20 ±0.17 and 0.36 ±0.12for the A_{2A}AR (**Table 4.2**). This indicates that the A_1AR and A_{2A}AR 100-fold selective compounds are sufficiently chemically distinct from each other to be correctly predicted by the model and that they show a clear structure-activity relationship. Conversely, dual and non-binders are suggested to share chemical similarities with both the A_1AR- and A_{2A}AR-selective classes, making it more challenging for the model to differentiate between these classes (**Figure 4.2**). Furthermore, the sensitivity and the positive predictive value (PPV) were consistently higher for A_{2A}AR-selective compounds than for A_1AR-selective compounds, whereas specificity and negative predictive value (NPV) were higher for A_1AR-selective compounds.

The non-binder class contains compounds that are inactive at both receptors. However, these compounds are not well differentiated from the active classes (A_1AR-, A_{2A}AR-selective, and dual), as is observed by low MCC (0.15 ±0.06) and poor ROC (0.57 ±0.07) for the non-binder class. The next section therefore describes bioactivity modeling of the A_1AR and A_{2A}AR in an attempt to categorize non-binders.

Table 4.2. Performance of selectivity classification models.

Classification model	Class	MCC	Sensitivity	Specificity	PPV	NPV	ROC
QSAR 2-class	A_1AR	0.40 ±0.13	0.62 ±0.16	0.76 ±0.11	0.57 ±0.12	0.86 ±0.07	0.87 ±0.06
	A_{2A}AR	0.40 ±0.13	0.76 ±0.11	0.62 ±0.16	0.86 ±0.07	0.57 ±0.12	0.87 ±0.06
QSAR 3-class	A_1AR	0.22 ±0.15	0.25 ±0.15	0.96 ±0.02	0.31 ±0.14	0.93 ±0.02	0.76 ±0.06
	A_{2A}AR	0.20 ±0.17	0.33 ±0.16	0.88 ±0.02	0.35 ±0.14	0.83 ±0.04	0.62 ±0.09
	Dual	0.10 ±0.09	0.81 ±0.03	0.29 ±0.10	0.75 ±0.02	0.35 ±0.08	0.58 ±0.06
QSAR 3-class	A_1AR	0.00 ±0.07	0.11 ±0.05	0.88 ±0.07	0.10 ±0.06	0.91 ±0.02	0.64 ±0.09
	A_{2A}AR	0.36 ±0.12	0.47 ±0.14	0.85 ±0.10	0.59 ±0.11	0.85 ±0.02	0.65 ±0.09
	Non-binder	0.07 ±0.13	0.67 ±0.11	0.39 ±0.12	0.70 ±0.05	0.37 ±0.12	0.50 ±0.09
QSAR 4-class	A_1AR	0.11 ±0.10	0.12 ±0.09	0.97 ±0.01	0.18 ±0.11	0.95 ±0.01	0.70 ±0.07
	A_{2A}AR	0.25 ±0.16	0.29 ±0.13	0.94 ±0.02	0.39 ±0.16	0.90 ±0.02	0.67 ±0.09
	Dual	0.09 ±0.05	0.51 ±0.05	0.58 ±0.08	0.50 ±0.05	0.60 ±0.02	0.58 ±0.05
	Non-binder	0.15 ±0.06	0.51 ±0.09	0.64 ±0.06	0.47 ±0.04	0.68 ±0.05	0.57 ±0.07

Means of 5-fold cross-validation with standard error of the mean (SEM). The class indicates the performance for that particular selectivity class: A_1AR-selective, A_{2A}AR-selective, dual (non-selective), and non-binders. MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, NPV = Negative Predictive Value, and ROC = receiver operating characteristic.

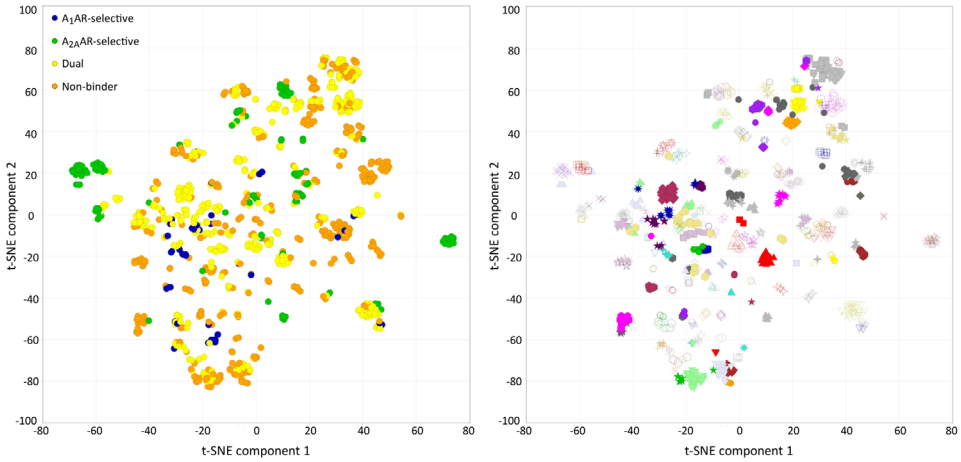


Figure 4.2. Chemical similarity of compounds of the selectivity classes A_1AR-selective, A_{2A}AR-selective, dual, and non-binders. The chemical similarity is visualized with t-SNE[20] based on FCFP4 fingerprints. (A) The used chemical clusters of the compounds: A_1AR-selective, A_{2A}AR-selective, dual binders, and non-binders. (B) Clusters based on chemical similarity; each color-symbol combination represents a unique cluster (136 clusters in total).

Modeling A_1AR and A_{2A}AR bioactivity using classification and regression QSAR models

The bioactivity of compounds for the A_1AR and A_{2A}AR were modeled with both classification and regression models. Classification models categorize compounds with using a pre-defined threshold (here pActivity ≥ 6.5) as 'active' and compounds below that threshold are termed 'inactive'. The model is trained on these activity classes and provides an activity class for test compounds as well. In contrast, regression models are not trained on classes, but on numerical bioactivity values. The output that is generated from a regression model is a bioactivity value, which can subsequently be assigned to an activity class. As can be observed in **Table 4.1** where the median pActivity for the sets is shown, this value (pActivity 6.5) is applicable for these data sets and was previously also shown to be a relevant threshold leading to balanced classes.[15]

Bioactivity classification and regression QSAR models were trained on the A_1AR/A_{2A}AR dataset, the same dataset that was used in the selectivity-classification QSAR models described in the previous section. Additionally, semi-selective compounds were added to increase the amount of training data. These semi-selective compounds have experimental activities for both receptors but do not fit into any of the four selectivity classes (e.g. a compound with pActivity A_1AR = 7.0, pActivity A_{2A}AR = 8.1). However, for bioactivity modeling the selectivity class is irrelevant, and thus these compounds were now included to increase model performance. Additionally, separate bioactivity QSAR models were trained on the A_1AR and A_{2A}AR bioactivity datasets. The validation test sets were composed based on chemical clusters and bioactivity of compounds; each subset contained both actives and inactives. These validation sets were not equal to the aforementioned (selectivity) validation sets as they were used for a different purpose: validation of bioactivity models instead of selectivity models. All bioactivity models were validated using the same cross-validation test sets, regardless of the dataset that was used in training (A_1AR/A_{2A}AR dataset, A_1AR bioactivity, or A_{2A}AR bioactivity). The A_1AR and A_{2A}AR

bioactivity datasets contained more data points than the $A_1AR/A_{2A}AR$ dataset as these sets also included single points (bioactivity measured only for one of the two receptors). The single bioactivity points were included in training, but excluded from validation to retain comparability of performance for the different models. Single points that belonged to the same chemical cluster as the data points in the test set were also excluded from training to prevent bias. The regression models show good model quality in training, with a high R^2 ($\geq$ 0.98) and low RMSE values ($\leq$ 0.14). Unfortunately, when applied on the validation set, performances are lower than expected based on training performance (likely caused by the challenging test set based on chemical clustering). Nevertheless judging the model performance on classification validation metrics a realistic estimation can be made for the predictive performance of the models.

The mean performances of all bioactivity models (based on 5-fold cross-validation) were better for the $A_{2A}AR$ than for the A_1AR, with an average ROC difference of 0.12 (**Table 4.3**). Furthermore, classification models performed worse than regression models, as indicated by their lower values for enrichment (ROC) and MCC: average difference in ROC for the A_1AR = 0.20 and $A_{2A}AR$ = 0.07, and average difference in MCC for the A_1AR = 0.07 and $A_{2A}AR$ = 0.03. Moreover, the MCC and ROC for A_1AR bioactivity classification models even indicated performances worse than random (MCC < 0 and ROC < 0.5). The best performing bioactivity models were based on regression, which reached an average performance (ROC score 0.60-0.70) for predicting bioactivities.

Modeling $A_1AR/A_{2A}AR$ subtype selectivity using regression models

The application of the above bioactivity model approach was tested in modeling the selectivity of compounds (i.e. modeling the affinity on the respective receptors and deriving selectivity from that indirectly). As the predicted bioactivities of the two bioactivity models are not correlated with each other, a separate validation was performed to indicate the performance of selectivity predictions. However, the cross-validation sets of the models in **Table 4.3** were clustered based on bioactivity instead of selectivity classes, hence bioactivity models were retrained using differently composed cross-validation sets to justify comparison with a later discussed selectivity-window model. Regression bioactivity models were selected as these outperformed the classification bioactivity models. Moreover, regression was preferred in selectivity modeling as with regression a quantitative value for selectivity can be derived.

Thus, bioactivity regression models were used to predict compound activity for the A_1AR and $A_{2A}AR$. Models were trained on the $A_1AR/A_{2A}AR$ dataset including additional 'semi-selective' compounds. The difference in predicted bioactivity for the two receptors was calculated as selectivity value. Subsequently, selectivity classes (A_1AR-selective, $A_{2A}AR$-selective, dual) were assigned to the compounds based on the predicted selectivity according to the same categorization rules as used in classification previously. The application of the two combined bioactivity models to deduce selectivity (two-step A_1AR-$A_{2A}AR$ difference model) resulted in models with average performance. Although ROC scores were decent (0.75 $\pm$0.09 and 0.72 $\pm$0.15), MCC was poor (0.19 $\pm$0.16 and 0.28* $\pm$0.12, *one failed validation) for both the A_1AR and $A_{2A}AR$, indicating that models were capable of ranking the compounds but less capable of explaining the whole data set (**Table 4.4**).

Table 4.3. Performances of classification and regression bioactivity models for the A_1AR and $A_{2A}AR$.

Protein	Model training type	Dataset in training	Validation set (only A_1AR or $A_{2A}AR$ compounds, respective of the protein)	MCC	Sensitivity	Specificity	PPV	NPV	ROC
A_1AR	Classification	A_1AR compounds in the $A_1AR/A_{2A}AR$ dataset + semi-selective compounds	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	-0.09 ±0.06	0.44 ±0.09	0.48 ±0.11	0.46 ±0.09	0.45 ±0.06	0.41 ±0.05
	Classification	A1AR bioactivity dataset	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	-0.16 ±0.05	0.39 ±0.08	0.45 ±0.10	0.42 ±0.09	0.41 ±0.06	0.39 ±0.04
	Regression	A_1AR compounds in the $A_1AR/A_{2A}AR$ dataset + semi-selective compounds	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	0.09 ±0.04	0.53 ±0.09	0.56 ±0.08	0.54 ±0.09	0.54 ±0.08	0.61 ±0.03
	Regression	A_1AR bioactivity dataset	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	0.04 ±0.06	0.46 ±0.08	0.58 ±0.08	0.52 ±0.10	0.52 ±0.08	0.59 ±0.05
	Regression	A_1AR bioactivity dataset	A_1AR bioactivity dataset	0.06 ±0.07	0.49 ±0.07	0.58 ±0.08	0.53 ±0.07	0.54 ±0.06	0.60 ±0.05
$A_{2A}AR$	Classification	$A_{2A}AR$ compounds in the $A_1AR/A_{2A}AR$ dataset + semi-selective compounds	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	0.11 ±0.09	0.59 ±0.10	0.50 ±0.13	0.73 ±0.05	0.39 ±0.05	0.59 ±0.06
	Classification	$A_{2A}AR$ bioactivity dataset	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	0.16 ±0.11	0.57 ±0.12	0.56 ±0.13	0.75 ±0.06	0.45 ±0.09	0.61 ±0.07
	Regression	$A_{2A}AR$ compounds in the $A_1AR/A_{2A}AR$ dataset + semi-selective compounds	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	0.12 ±0.10	0.69 ±0.10	0.40 ±0.08	0.70 ±0.04	0.47 ±0.10	0.64 ±0.06
	Regression	$A_{2A}AR$ bioactivity dataset	$A_1AR/A_{2A}AR$ dataset + semi-selective compounds	0.21 ±0.07	0.64 ±0.10	0.56 ±0.10	0.76 ±0.04	0.46 ±0.04	0.69 ±0.05
	Regression	$A_{2A}AR$ bioactivity dataset	$A_{2A}AR$ bioactivity dataset	0.19 ±0.07	0.63 ±0.11	0.54 ±0.09	0.70 ±0.04	0.50 ±0.05	0.69 ±0.05

Query compounds were categorized based on post-classification of the bioactivity predictions: predicted pActivity < 6.5 = inactive and predicted pActivity ≥ 6.5 = active. MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, NPV = Negative Predictive Value, and ROC = receiver operating characteristic.

Table 4.4. Performances of selectivity modeling using the two-step A_1AR-$A_{2A}AR$ difference model or the selectivity-window model.

Model	Class	MCC	Sensitivity	Specificity	PPV	NPV	ROC
Trained on all double points, tested on all double points (non-binders and semi-selective compounds always true/false negative)							
A1AR-A2AAR difference	A_1AR	0.15 ±0.13	0.17 ±0.14	0.99 ±0.00	0.18 ±0.12	0.98 ±0.01	0.76 ±0.09
	$A_{2A}AR$	0.11 ±0.07	0.19 ±0.09	0.94 ±0.02	0.15 ±0.07	0.94 ±0.01	0.67 ±0.14
	Dual	0.26 ±0.05	0.76 ±0.02	0.50 ±0.07	0.68 ±0.02	0.59 ±0.03	0.66 ±0.02
Selectivity-window	A_1AR	0.07*** ±0.09	0.03 ±0.03	0.99 ±0.00	0.10*** ±0.10	0.97 ±0.00	0.87 ±0.03
	$A_{2A}AR$	0.22 ±0.12	0.15 ±0.09	0.99 ±0.00	0.42 ±0.18	0.94 ±0.01	0.74 ±0.07
	Dual	0.36 ±0.06	0.85 ±0.03	0.48 ±0.03	0.69 ±0.01	0.70 ±0.06	0.70 ±0.02
Trained on all double points, tested on only A_1AR, $A_{2A}AR$, and dual class							
A1AR-A2AAR difference	A_1AR	0.19 ±0.16	0.17 ±0.14	0.99 ±0.00	0.28 ±0.18	0.96 ±0.01	0.75 ±0.09
	$A_{2A}AR$	0.28* ±0.12	0.19 ±0.09	0.98 ±0.01	0.48* ±0.15	0.91 ±0.01	0.72 ±0.15
	Dual	0.25 ±0.07	0.76 ±0.02	0.57 ±0.11	0.91 ±0.02	0.28 ±0.04	0.70 ±0.05
Selectivity-window	A_1AR	0.17*** ±0.23	0.03 ±0.03	0.99 ±0.01	0.50*** ±0.50	0.92 ±0.01	0.88 ±0.04
	$A_{2A}AR$	0.32* ±0.15	0.15 ±0.09	1.00 ±0.00	0.75* ±0.25	0.82 ±0.03	0.80 ±0.07
	Dual	0.33 ±0.11	0.84 ±0.04	0.46 ±0.06	0.80 ±0.02	0.55 ±0.10	0.66 ±0.04
Trained on only A_1AR, $A_{2A}AR$, and dual class, tested on only A_1AR, $A_{2A}AR$, and dual class							
Selectivity-window	A_1AR	-0.05**** ±0.00	0.00 ±0.00	0.99 ±0.01	0.00**** ±0.00	0.92 ±0.01	0.81 ±0.05
	$A_{2A}AR$	0.23 ±0.18	0.25 ±0.12	0.94 ±0.03	0.43 ±0.22	0.83 ±0.03	0.66 ±0.11
	Dual	0.04 ±0.12	0.69 ±0.05	0.35 ±0.09	0.73 ±0.03	0.32 ±0.09	0.53 ±0.08

The class indicates the performance for that particular selectivity class: A_1AR-selective, $A_{2A}AR$-selective, and dual (non-selective). The query compounds were categorized based on post-classification of the selectivity predictions: A_1AR-selective when pActivity difference ≥ 2, $A_{2A}AR$-selective when pActivity difference ≤ -2, and dual binder when pActivity difference ≥ -1 & ≤ 1. *1/5 folds failed, **2/5 folds failed, ***3/5 folds failed, ****4/5 folds failed, MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, NPV = Negative Predictive Value, and ROC = receiver operating characteristic.

Continuing our single model approach to predict selectivity we explored the usage of regression for a selectivity-window model rather than classification. In contrast to the two-step A_1AR-$A_{2A}AR$ difference model (regression), the single model to predict selectivity between the A_1AR and $A_{2A}AR$ was based on the difference in affinity rather than the prediction of bioactivity and calculation of the resulting selectivity. The regression model was trained directly on the difference in bioactivity for both receptors (pActivity A_1AR – pActivity $A_{2A}AR$ = selectivity-window) and predicts a quantitative score for the degree of selectivity of a compound (difference in pActivity). A positive score indicates A_1AR-selectivity, a negative score $A_{2A}AR$-selectivity, and a score close to zero indicates dual binders. The model was evaluated based on the ROC score and classification metrics (MCC, sensitivity, specificity, PPV, and NPV). The rules for classification of the A_1AR-, $A_{2A}AR$-selective, and dual binders were derived from the thresholds applied in the selectivity classification models: $A_1AR \geq 100$ fold selective equals pActivity difference ≥ 2, $A_{2A}AR \geq 100$ fold selective equals pActivity difference ≤ -2, and for dual binders (≤ 10 fold selective) pActivity difference ≥ -1 & ≤ 1.

The selectivity-window regression model was trained on the same data ($A_1AR/A_{2A}AR$ dataset and semi-selective compounds) as the two-step A_1AR-$A_{2A}AR$ difference model described above in which selectivity was deducted from two separate bioactivity models. The selectivity-window outperformed the two-step A_1AR-$A_{2A}AR$ difference model with increased ROC values for selectivity classes A_1AR- and $A_{2A}AR$-selective (ROC increase 0.07-0.13) (**Table 4.4**).

Figure 4.3 shows example compounds that were misclassified with the two-step A_1AR-$A_{2A}AR$ difference model, but were correctly predicted using the selectivity-window model. The similarity (Tanimoto FCFP4) between the mispredicted compounds by the two-step A_1AR-$A_{2A}AR$ difference model was 0.25, whereas the similarity within wrongly predicted compounds by the selectivity-window model was 0.58. This indicates that the two-step A_1AR-$A_{2A}AR$ difference model is challenged by selectivity prediction of more diverse compounds and the selectivity-window model underperforms on specific chemical scaffolds. The most frequently mispredicted scaffold by the selectivity-window model was N-(2-(furan-2-yl)-6-(1H-pyrazol-1-yl)pyrimidin-4-yl)-2-phenoxyacetamide (**Figure 4.4**).

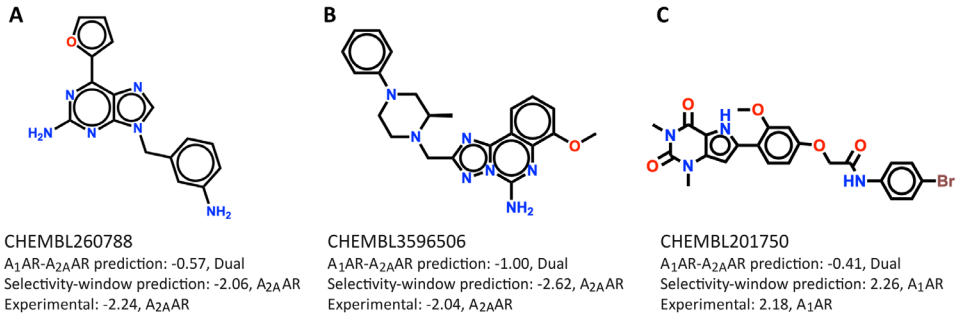
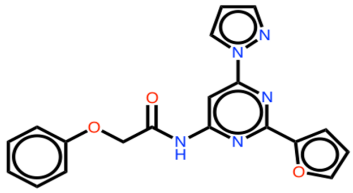


Figure 4.3. Chemical structures of compounds with predictions by different selectivity models. The compounds were wrongly predicted with the two-step A_1AR-$A_{2A}AR$ model and correctly predicted with the selectivity-window model. Predictions are indicated as: predicted A_1AR-selective (A_1AR), $A_{2A}AR$-selective ($A_{2A}AR$), and as dual binder (Dual) for ligands CHEMBL260788 (A), CHEMBL3596506 (B), and CHEMBL201750 (C).



N-(2-(furan-2-yl)-6-(1H-pyrazol-1-yl) pyrimidin-4-yl)-2-phenoxyacetamide

Figure 4.4. Most frequent chemical scaffold of compounds that were wrongly predicted by the selectivity-window model, but correctly predicted by the two-step A_1AR-$A_{2A}AR$ difference model.

It should be noted that for the A_1AR and $A_{2A}AR$ predictions MCC and PPV could not always be calculated in cross-validation, resulting in failed cross-validation folds, or iterations. This is explained by the lack of true/false positives in the particular cross-validation fold: PPV cannot be calculated if there are no positives, MCC cannot be calculated without a PPV.

The selectivity-window model performed better than the three-class selectivity classification model validated earlier (**Table 4.2**). It was tested whether this increased performance was a result from the increased amount of data as the selectivity-window model included additional semi-selective compounds and non-binders, which the three-class selectivity classification model necessarily lacked as the affinity difference was not large enough to meet the classification cut-off. When these additional data points were excluded from the selectivity-window model, the ROC dropped from 0.78 to 0.67 (average over classes), which is comparable with the ROC of 0.65 of the three-class selectivity classification model. This observation clearly confirms a direct link between model quality and data availability and shows that the increased performance of the selectivity-window model is attributed to additional data points. Hence it is advantageous to use continuous models in selectivity modeling as in this case more data can be included. In addition to the benefit of increased data availability, continuous selectivity models also provide the ability to calculate a selectivity ratio as opposed to the class only. This selectivity ratio indicates the degree of selectivity and therefore cannot only identify selective compounds, but can also differentiate highly selective compounds from weakly selective compounds.

Remarkably, metrics based on classification (MCC, sensitivity, specificity, PPV, and NPV) for the selectivity-window model (without non-binders and semi-selective compounds) (**Table 4.4**) are lower for the A_1AR-selective compounds and dual binders than metrics of the 3-class classification model (both trained on the same data) (**Table 4.2**), whereas ROC scores are comparable or higher. Therefore, the predictions of the selectivity-window model were compared with the experimentally measured selectivity values (**Figure 4.5**). It was observed that the A_1AR-selective compounds have consistently lower selectivity-window predictions than the experimental selectivity values. As a result, fewer compounds reached the A_1AR-selective classification threshold, decreasing the number of the A_1AR-selective positives drastically. From the 50 A_1AR-selective compounds, none reached the threshold. Of all predictions, only three compounds reached the A_1AR-selective threshold, which were dual binders instead of the A1AR-selective compounds. This deficiency of predicted positives explains the failed cross-validation calculations for the A_1AR-selective class.

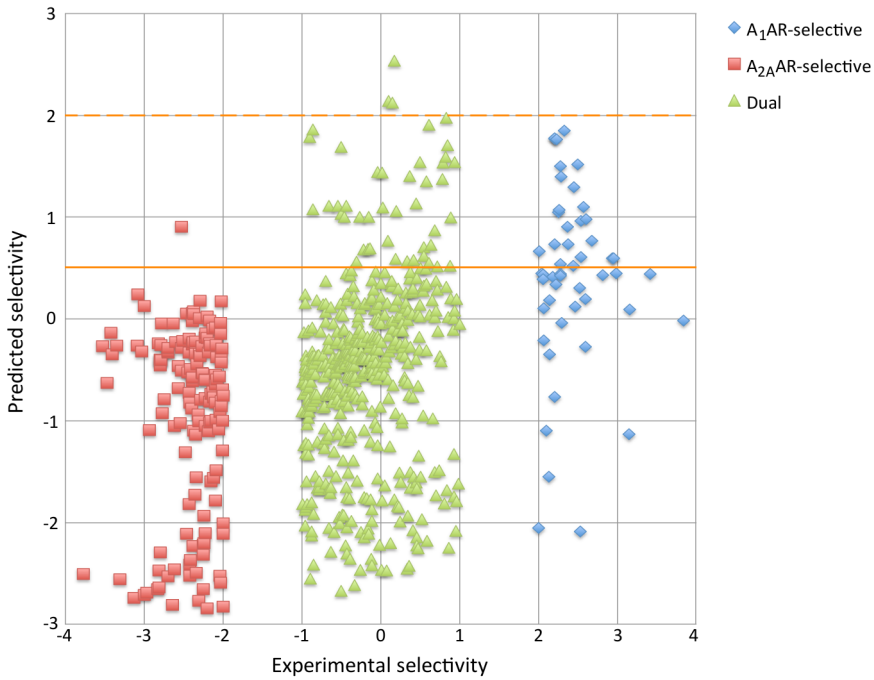


Figure 4.5. Relationship between experimental selectivity versus predicted selectivity. Predicted selectivity values shown for the selectivity-window model. A_1AR-selective classification thresholds shown as orange lines (dotted = old threshold, solid = new threshold).

To compensate, classification validation metrics were re-calculated post-hoc using classification thresholds that were adapted to compensate for the generalization of selectivity for A_1AR-selective compounds. Compounds were deemed A_1AR-selective when the selectivity-window ≥ 0.5, A_{2A}AR-selective when the selectivity-window ≤ -2 (unchanged), and dual binder when the selectivity-window ≥ -1 & < 0.5. Using the new thresholds, values of the metrics for the A_1AR-selectivity predictions improved: MCC 0.31 ±0.10, sensitivity 0.45 ±0.12, specificity 0.91 ±0.02, PPV 0.32 ±0.06, and NPV 0.95 ±0.01, indicating that the revised threshold improves the categorization of the A_1AR-selective compounds. Not all A_1AR-selective compounds were correctly categorized but the post-hoc optimized threshold was considered adequate, as lowering the A_1AR-selective threshold further would increase sensitivity (by categorization of more compounds as A_1AR-selective), but would also decrease PPV. Here, the correctness of predictions was prioritized over the number of predicted active compounds; hence PPV was prioritized over sensitivity.

Removal of non-binders to enhance performance

Although the selectivity-window model differentiates between the A_1AR-, A_{2A}AR-selective compounds, and dual binders, the model does not consider potential inactivity of compounds. Consequently, non-binders cannot be filtered using this model. Therefore, a consensus approach of statistical modeling and structure-based docking was applied to identify and exclude non-binders.

Bioactivity regression models described above for the A_1AR and $A_{2A}AR$ were combined with docking of the $A_1AR/A_{2A}AR$ dataset and semi-selective compounds into crystal structures of both proteins. Bioactivity predictions for the A_1AR and $A_{2A}AR$, and selectivity-window predictions, were derived for the entire $A_1AR/A_{2A}AR$ dataset and semi-selective compounds by assembling the predictions made during 5-fold cross-validation of the previously trained regression models. Compounds were docked into crystal structures of the A_1AR (PDB: 5UEN)[10] and the $A_{2A}AR$ (PDB: 5OLZ)[21], which resulted in docking scores for both receptors. Compounds were assigned a separate bioactivity label for the A_1AR and $A_{2A}AR$: compounds in the A_1AR were labeled 'active' when predicted pActivity ≥ 7 and docking score ≤ -9. Compounds in the $A_{2A}AR$ were labeled 'active' when predicted pActivity ≥ 7 and docking score ≤ -10.

Compounds with predicted selectivity-window ≥ 0.5 or ≤ -2, corresponding with A_1AR- and $A_{2A}AR$-selective, were subsequently filtered using the consensus bioactivity filter (**Figure 4.6**). The PPV for A_1AR-selective compounds drastically increases from 0.13 to 0.39 when the selectivity-window predictions were filtered using the consensus bioactivity filter for the A_1AR. The $A_{2A}AR$ bioactivity filtering also increased the PPV of $A_{2A}AR$-selective compounds; here docking and consensus filtering performed equally well (PPV $A_{2A}AR$-selective: 0.80).

Non-binders that were not removed using the statistical filter only, but were filtered when the consensus approach was used, were inspected in the crystal structure of the A_1AR. Some non-binders (e.g., CHEMBL1800792) did not adapt a favorable conformation when docked into the A_1AR (**Figure 4.7**). Moreover, an interaction with pocket residue $Asn^{6.55}$ (Ballesteros-Weinstein numbering) was frequently not observed. This is an essential missing element, as interaction with this residue has shown to be important for ligand binding to the A_1AR and $A_{2A}AR$.[11,22] However, some non-binders were able to make this interaction (CHEMBL372580), but nevertheless had

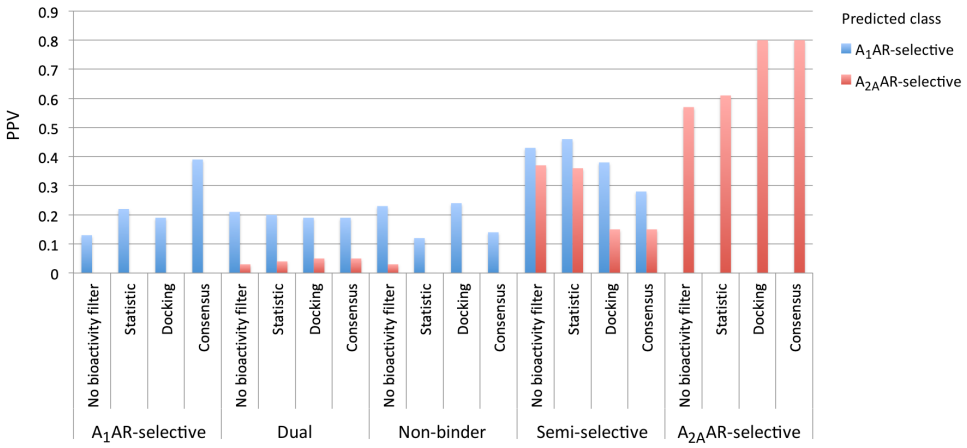


Figure 4.6. Positive predictive value (PPV) of compounds predicted to be A_1AR- or $A_{2A}AR$-selective. The PPV depicts the number of experimentally validated selective compounds divided by the total number of predicted selective compounds. PPVs are shown when different filters are applied: no bioactivity filter, statistical bioactivity, bioactivity based on docking score, and consensus bioactivity (statistical bioactivity and structure-based docking).

a docking score that did not reach the set threshold (docking score ≤ -9). The poses of the non-binders were compared to those of an A_1AR-selective (CHEMBL207824) and an A_{2A}AR-selective compound (CHEMBL371436). Both selective compounds adapt a conformation that is able to make an interaction with residue $Asn^{6.55}$. Furthermore, the poses also constitute the same aromatic interactions (with $Phe171^{EL2}$ in the A_1AR and with $Phe168^{EL2}$ and $His250^{6.52}$ in the A_{2A}AR) as the co-crystalized ligands and adapt a similar scaffold orientation. Finally, the poses of the selective ligands have favorable docking scores (-10.30 and -10.91, respectively), supporting that these compounds are binders for the A_1AR or A_{2A}AR.

Validation of the selectivity-window model on an external set

The predictive selectivity-window model (trained on A_1AR/A_{2A}AR dataset and semi-selective compounds) was challenged to predict the selectivity of compounds from an external validation set. This set contained 1,482 compounds of which a dose-response bioactivity value (K_i/IC_{50}/EC_{50}/K_d) was known for at least one of the two receptors. If an accurate bioactivity value was available for both receptors, the compound was classified according to prior rules applied in this study. However, if an accurate bioactivity value for only one receptor was known, a less accurate bioactivity measurement (inhibition as percentage displacement/efficacy/change) was used to identify inactivity for the missing receptor. The low accuracy of the bioactivity values makes this data less suitable for model training on the quantitative difference between activity on the two receptors, but suitable for classification validation. A pChEMBL value of < 4.5 or inhibition threshold of ≤ 50% (at 10 μM) was used to label inactive compounds, whereas a pChEMBL value of ≥ 6.5 was used to indicate active compounds. The selectivity-window model was applied to the compounds in the external validation set, providing them all with a predicted selectivity score and, subsequently, a selectivity class. The validation encompassed all selectivity classes: A_1AR- and A_{2A}AR-selective compounds, dual binders and non-binders. Since the selectivity-window model has no threshold for non-binders, non-binders were always considered either as true or false negative (never false/true positive).

Without filtering inactives, the selectivity-window model performed average in the prediction of the A_1AR-selective compounds (ROC 0.75) and A_{2A}AR-selective compounds (ROC 0.66) in the external validation set (**Table 4.5**). However, application of the consensus bioactivity filter resulted in an increase of the classification enrichment of the A_1AR- and A_{2A}AR-selective compounds. Although the ROC for A_1AR-selective compounds decreased after applying the bioactivity filter, PPV, and thus the fraction of true A_1AR-selective compounds compared to all predicted A_1AR compounds, increased from 0.12 to 0.21. In addition, MCC increased slightly from 0.13 to 0.18. Inspection of the compounds showed that all non-binders were removed after filtering the selectivity-window predictions with the consensus bioactivity filter. The decrease in ROC for the A_1AR-selective class was thus caused by the presence of dual binders only. Remarkably, sensitivity for the A_1AR- and A_{2A}AR-selective compounds was both 1.00 (100%), whereas sensitivity for dual binders was 0.00 (0%). Although dual compounds were present in the set that was filtered with the selectivity-window model, these compounds were wrongly categorized as either A_1AR- or A_{2A}AR-selective. The predicted dual binders prior to bioactivity filtering were, in fact, non-binders. However, these non-binders were correctly filtered out using the bioactivity filter, leaving the dual binder class without positive-predicted compounds. Note that the results do not specify dual binder enrichment, as the bioactivity predictions encompassed only compounds predicted to be A_1AR- or A_{2A}AR-selective.

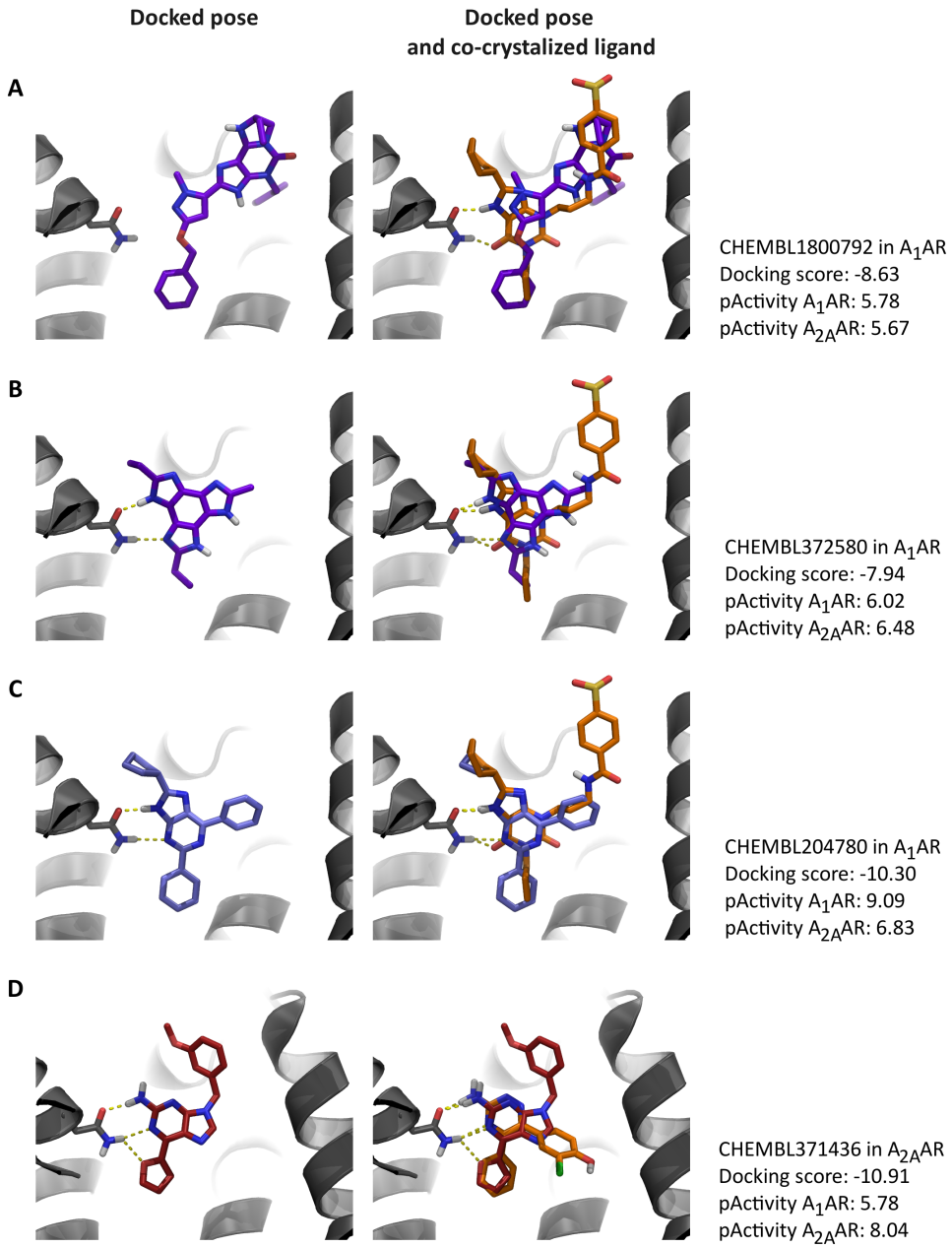


Figure 4.7. Docked poses of compounds in their corresponding targets. Poses of two non-binders in the A_1AR (CHEMBL1800792 in (A) and CHEMBL372580 in (B)), an A_1AR-selective compound (CHEMBL204780 in (C)), and $A_{2A}AR$-selective compound (CHEMBL371436 in (D)). Docked poses are compared to the co-crystalized ligands shown in orange. Hydrogen bonds between ligands and $Asn^{6.55}$ are shown in yellow.

Table 4.5. Performance of the selectivity-window model on an external validation set.

Model	Class	MCC	Sensitivity	Specificity	PPV	NPV	ROC
Selectivity-window	A_1AR	0.13	0.39	0.83	0.12	0.96	0.75
	$A_{2A}AR$	0.40	0.24	1.00	0.70	0.97	0.66
	Dual	0.02	0.81	0.21	0.64	0.39	0.37
Selectivity-window and bioactivity filtered	A_1AR	0.18	1.00	0.16	0.21	1.00	0.66
	$A_{2A}AR$	0.88	1.00	0.97	0.80	1.00	0.98
	Dual	-	0.00	1.00	-	0.28	0.72

The query compounds were categorized based on post-hoc optimized classification of the selectivity predictions: A_1AR-selective when selectivity-window ≥ 0.5, $A_{2A}AR$-selective when selectivity-window ≤ -2, and dual binder when selectivity-window ≥ -1 & < 0.5. MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, NPV = Negative Predictive Value, and ROC = receiver operating characteristic.

Discussion

While QSAR models are widely applied in bioactivity modeling, they can also effectively be used in selectivity modeling. However, modeling of selectivity requires a substantial amount of data, as activities for more than one protein have to be measured. The amount of data that is available influences the performance of the selectivity model as was observed for the performances of the selectivity-window models when trained on limited data. To increase the amount of data that is sufficient for selectivity modeling continuous regression models can be applied instead of classification models. With regression not only compounds that belong to a defined selectivity class can be included, but also compounds of which there is some selectivity but not large enough to fit into a class. Another benefit of regression is that the degree of selectivity can be provided in addition to the selectivity class of a compound.

Multiple QSAR regression models to derive selectivity for a panel of kinases were used by Sciabola et al.[23]. First, regression bioactivity models were trained for every kinase in the panel. Next, bioactivity patterns were predicted for a set of compounds against all kinases, from which subsequently selectivity was derived. To compare, we repeated a similar approach was repeated by us in the current work. However, we also introduce the selectivity-window model, which is a direct implementation of selectivity. We show that this approach outperformed models that predicted selectivity indirectly by using separate bioactivity models. Even though separate bioactivity models can include more data since compounds measured for just one protein can be considered, this approach did not increase model performance enough to outperform the selectivity-window model.

Nevertheless, an advantage of using separate bioactivity models to deduce selectivity from is that additional proteins can be added easily: the selectivity between the added protein and existing proteins can quickly be deduced from the results of the added bioactivity model. In selectivity-window modeling multiple models need to be trained to predict selectivities of compounds against a panel of targets: one model for every target-target combination. However, while using separate bioactivity models can be more convenient, selectivity-window modeling may yield more accurate predictions. It should also be noted that automatically generating these models using scripting can be considered trivial. Therefore, when sufficient selectivity data is available it is worthwhile to apply selectivity-window modeling.

The higher accuracy of selectivity-window modeling compared to using multiple bioactivity models is suggested to be influenced by the higher quality of the data used in selectivity-window modeling. In selectivity-window modeling, selectivity is predicted based on data from biological experiments. Although biological experiments are susceptible to error (on average an error of 0.6 log units[24]), this data is more reliable than data derived from statistical models whose error by definition should be higher than the error of the data they were trained on. In practice the error of statistical models (and hence the error of predictions) varies around 0.5-1.0 log units.[25,26] While this error may accumulate with the experimental error, there is also the possibility that modeling can reduce part of the experimental error. An additional study is required to reveal how the modeling error behaves in combination with the experimental error.

The selectivity-window model by itself is not capable of distinguishing actives from inactives as it is trained on the difference only, which is different from the affinity. However, separate bioactivity models can be applied to filter potentially selective compounds from inactives, or non-binders. A study by Zhao et al., where subtype selectivity between epigenetic targets HDAC1 and HDAC6 was modeled by classification of selectivity, utilizes a comparable approach by first predicting selectivity, followed by bioactivity.[27] However, the selectivity model in that study is incapable of predicting the degree of selectivity as a classification model was used. Furthermore, only statistical models were used by the authors to predict bioactivity of compounds. In the current study it was observed that implementation of statistical bioactivity models only increased the enrichment of selective compounds slightly. In contrast, addition of structure-based docking scores increased the enrichment of selective compounds substantially for both the A_1AR and A_{2A}AR. Moreover, structure-based docking performed equally well as the consensus model (statistical bioactivity and docking) for A_{2A}AR-selective enrichment.

Conclusion

We demonstrated that continuous QSAR models can be applied to model selectivity on the A_1AR and A_{2A}AR. The selectivity-window model, which was trained directly on the difference in affinity between both receptors, outperformed a two-step A_1AR-A_{2A}AR selectivity model. In the two-step model, which is generally applied in selectivity modeling, selectivity predictions are derived indirectly by calculation of the difference between bioactivity predictions that resulted from two separate models. Even though the separate bioactivity models included more data, the performance did not increase enough to outperform our selectivity-window model. Furthermore, a combination of statistical bioactivity models and structure-based docking contributed to the enrichment of selective compounds and can be used to exclude non-binders (which are not predicted accurately when directly predicting selectivity). In summary, we demonstrated that accurate selectivity predictions can be made for the A_1 and A_{2A} adenosine receptors by combining the selectivity-window model and consensus bioactivity modeling. This method can easily be applied to other protein targets (e.g., kinases) as well, provided sufficient data is available.

Methods

Training/test datasets

The dataset was compiled from publicly available data derived from the ChEMBL database (Gaulton et al. 2012; Bento et al. 2014) (version 23) and in-house data from Leiden University (Leiden, The Netherlands). Compounds with experimental activities were collected for the human A_1AR (P30542) and human $A_{2A}AR$ (P29274). The data derived from ChEMBL was filtered on confidence score 7 and 9, and a pChEMBL value ≥ 4. In-house data was filtered similarly: activity ($K_i/IC_{50}/EC_{50}$) ≤ 10^{-4} M. K_i values were prioritized over IC_{50} or EC_{50} values. Thus, for duplicates, when more than one type was available for a given compound-receptor pair, K_i values were kept and IC_{50} and EC_{50} values were discarded. The mean value was taken when multiple bioactivity values of the same type were reported for a given compound-receptor pair (e.g., mean of multiple K_i values for the same compound). The standardized activity values are reported as pActivity values. An antagonist-focused dataset was compiled from the filtered data by removing compounds with a ribose or dicyanopyridine scaffold. From this antagonist-focused dataset an $A_1AR/A_{2A}AR$ dataset that contained only compounds with activities measured on both the A_1AR and $A_{2A}AR$ was derived. The compounds were assigned to the $A_1AR/A_{2A}AR$ dataset after they were categorized into one of the following five classes: non-binders when pActivity for both the A_1AR and $A_{2A}AR$ < 6.5, A_1AR-selective (pActivity ≥ 6.5 for the A_1AR and activity compared to the $A_{2A}AR$ ≥ 100-fold), $A_{2A}AR$-selective (pActivity ≥ 6.5 for the $A_{2A}AR$ and activity compared to the A_1AR ≥ 100-fold), and dual binders (pActivity ≥ 6.5 for both the $A_1AR/A_{2A}AR$ and activity difference ≤ 10-fold). Compounds with experimental bioactivities for both the A_1AR and $A_{2A}AR$, but that did not fit into any of the classes of the $A_1AR/A_{2A}AR$ dataset, were termed "semi-selective" compounds (855 bioactivities). The antagonist-focused dataset contained 5,897 activities, the $A_1AR/A_{2A}AR$ dataset included 1,106 compounds, of which 50 A_1AR-selective and 146 $A_{2A}AR$-selective. Additionally, the antagonist-focused dataset was split into two datasets for bioactivity modeling: the A_1AR bioactivity dataset (2,774 compounds) and the $A_{2A}AR$ bioactivity dataset (3,123 compounds).

T-distributed Stochastic Neighbor Embedding (t-SNE)

The chemical similarity of the $A_1AR/A_{2A}AR$ dataset was plotted using t-SNE.[20] Compounds were described using FCFP4 fingerprints (fixed-length array of bits 2024). Two dimensions were calculated: t-SNE component 1 and t-SNE component 2. The settings were as follows: maximum number of iterations 5000, theta 0, perplexity 30, momentum 0.5, final momentum 0.8, and learning rate 10. Additionally a t-SNE was conducted showing the distribution of chemically clustered compounds using affinity propagation (FCFP4).[28]

External validation set

An external validation set was created by using compounds that had been newly added in ChEMBL version 24 and 25. Furthermore, compounds with confidence score 6 and 8 from previous ChEMBL versions were added. Additionally, less accurate bioactivity measurements (e.g., % displacement) were used to identify inactives. These less accurate bioactivities included bioactivities measured as percentage displacement, efficacy and change. If a pChEMBL value was known for both receptors, the compounds were categorized into the selectivity classes A_1AR-selective, $A_{2A}AR$-selective, dual binder, and non-binder according to the same rules as

used for the $A_1AR/A_{2A}AR$ dataset. If a pChEMBL value was known for only one of the two receptors, less accurate measurements (displacement/efficacy/change) were used to identify if the compound was marked as inactive for the other receptor. Subsequently, a selectivity class could be assigned. A pChEMBL value of < 4.5 or inhibition threshold of $\leq 50\%$ (at 10 μM) was used to label compounds as inactive and a pChEMBL value of ≥ 6.5 was used to identify active compounds. Subsequently the selectivity class was derived from these bioactivity classes: A_1AR-selective if active on the A_1AR and inactive on the $A_{2A}AR$, $A_{2A}AR$-selective if inactive on the A_1AR and active on the $A_{2A}AR$, and non-binder if inactive on both the A_1AR and $A_{2A}AR$. Again, compounds with ribose and dicyanopyridine scaffolds were excluded, resulting in an external validation set of 1,482 compounds.

Machine learning

QSAR bioactivity and selectivity models were constructed using the R XGBoost model component in Pipeline Pilot (version 18.1.0.1604).[29] The following settings were applied for both classification and continuous models: maximum number of trees 1000, learning rate 0.3, maximum depth 7, data fraction 1.0, and descriptor fraction 0.7. Compound descriptors were calculated within the component and included ALogP, molecular weight, number of H-donors, number of H-acceptors, number of rotatable bonds, number of atoms, number of (aromatic) rings, and FCFP6 fingerprints (fixed-length array of bits 2024).

Cross-validation

The models were validated with 5-fold cross-validation where they were trained on 80% and tested on 20% of the dataset. The $A_1AR/A_{2A}AR$ dataset was split into five subsets that each contained all four classes (A_1AR-selective/$A_{2A}AR$-selective/dual/non-binder) and the A_1AR bioactivity dataset and $A_{2A}AR$ bioactivity dataset were both separated into five subsets considering an active/inactive distribution. This consideration of class-distribution ensured that every subset contained each (bioactivity) class, which allows for balanced model training and validation. Chemical similarity of compounds was also considered; the $A_1AR/A_{2A}AR$ dataset and bioactivity datasets were each split into five subsets with every set covering different chemical structures. In order to create five chemically distinct subsets, each selectivity/activity class was clustered into ten clusters with the cluster molecules component in Pipeline Pilot (based on FCFP4). Subsequently, the smallest and largest clusters were combined into one group. This was done recurrently until all clusters were divided into five groups with every group containing 2 clusters per class. Finally, the resulting groups of each selectivity/activity class were distributed equally, resulting in five chemically distinct subsets comprising all selectivity/activity classes. The model performances were evaluated using the following metrics: MCC, Matthews Correlation Coefficient; sensitivity; specificity; PPV, Positive Predictive Value; NPV, Negative Predictive Value; and ROC, receiver operating characteristic.[31,32] The (traditionally classification) metrics MCC, sensitivity, specificity, PPV, and NPV, were either derived from the classification models directly, or calculated from the output of the regression models.

Protein preparation and docking

Protein crystal structures were prepared with the protein prep wizard in Maestro 11, Schrödinger Suites 2017-4.[33] First, modified amino acid residues were mutated back to wild type. Next, the protein was prepared by filling in missing side chains, adding hydrogens, and creation of disulfide

bonds. Compounds were prepared for docking using LigPrep from Schrödinger Suites 2017-4. Different tautomers were generated and compound charges were calculated at pH 7.4. Docking was performed with Glide from Maestro 11, Schrödinger Suites 2017-4. SP (standard precision) was used in docking and 10 poses per compound were generated.

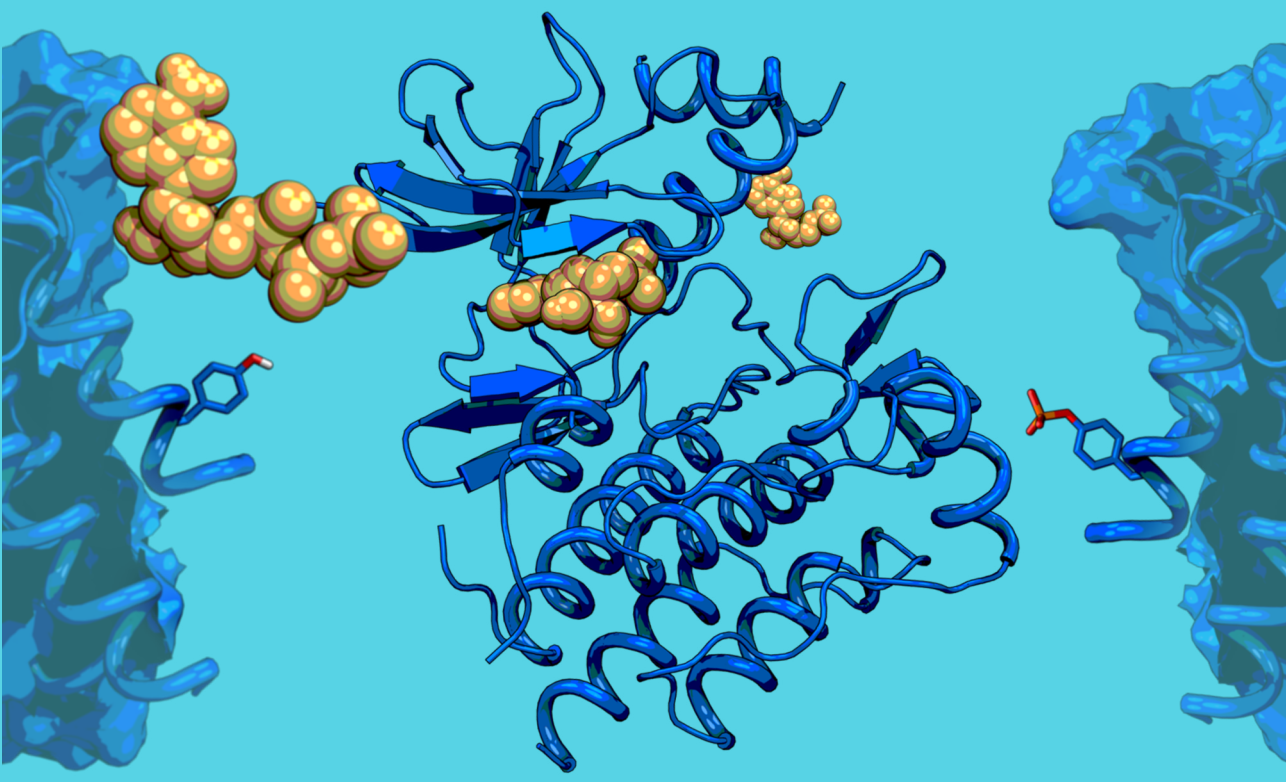
References

(1) Burggraaff, L.; Oranje, P.; Gouka, R.; van der Pijl, P.; Geldof, M.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 Using Proteochemometric Modeling. J. Cheminform. 2019, 11, 15. https://doi.org/10.1186/s13321-019-0337-8.
(2) van Westen, G. J. P.; Wegner, J. K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. PLoS One 2011, 6, e27518. https://doi.org/10.1371/journal.pone.0027518.
(3) Lenselink, E. B.; Jespers, W.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. Interacting with GPCRs: Using Interaction Fingerprints for Virtual Screening. J. Chem. Inf. Model. 2016, 56, 2053–2060. https://doi.org/10.1021/acs.jcim.6b00314.
(4) Lenselink, E. B.; Louvel, J.; Forti, A. F.; van Veldhoven, J. P. D.; de Vries, H.; Mulder-Krieger, T.; McRobb, F. M.; Negri, A.; Goose, J.; Abel, R.; van Vlijmen, H. W. T.; Wang, L.; Harder, E.; Sherman, W.; IJzerman, A. P.; Beuming, T. Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. ACS Omega 2016, 1, 293–304. https://doi.org/10.1021/acsomega.6b00086.
(5) Jespers, W.; Esguerra, M.; Åqvist, J.; Gutiérrez-de-Terán, H. QligFEP: An Automated Workflow for Small Molecule Free Energy Calculations in Q. J. Cheminform. 2019, 11, 26. https://doi.org/10.1186/s13321-019-0348-5.
(6) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. J. Med. Chem. 2014, 57, 7874–7887. https://doi.org/10.1021/jm5006463.
(7) Jacobson, K. A.; Gao, Z.-G. Adenosine Receptors as Therapeutic Targets. Nat. Rev. Drug Discov. 2006, 5, 247–264. https://doi.org/10.1038/nrd1983.
(8) Fredholm, B. B.; IJzerman, A. P.; Jacobson, K. A.; Klotz, K.-N.; Linden, J. International Union of Pharmacology. XXV. Nomenclature and Classification of Adenosine Receptors. Pharmacol. Rev. 2001, 53, 527 LP – 552.
(9) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. J. Mol. Biol. 1977, 112, 535–542. https://doi.org/10.1016/S0022-2836(77)80200-3.
(10) Glukhova, A.; Thal, D. M.; Nguyen, A. T.; Vecchio, E. A.; Jörg, M.; Scammells, P. J.; May, L. T.; Sexton, P. M.; Christopoulos, A. Structure of the Adenosine A1 Receptor Reveals the Basis for Subtype Selectivity. Cell 2017, 168, 867-877.e13. https://doi.org/10.1016/j.cell.2017.01.042.
(11) Cheng, R. K. Y.; Segala, E.; Robertson, N.; Deflorian, F.; Doré, A. S.; Errey, J. C.; Fiez-Vandal, C.; Marshall, F. H.; Cooke, R. M. Structures of Human A1 and A2A Adenosine Receptors with Xanthines Reveal Determinants of Selectivity. Structure 2017, 25, 1275-1285.e4. https://doi.org/https://doi.org/10.1016/j.str.2017.06.012.
(12) Draper-Joyce, C. J.; Khoshouei, M.; Thal, D. M.; Liang, Y.-L.; Nguyen, A. T. N.; Furness, S. G. B.; Venugopal, H.; Baltos, J.-A.; Plitzko, J. M.; Danev, R.; Baumeister, W.; May, L. T.; Wootten, D.; Sexton, P. M.; Glukhova, A.; Christopoulos, A. Structure of the Adenosine-Bound Human Adenosine A1 Receptor–Gi Complex. Nature 2018, 558, 559–563. https://doi.org/10.1038/s41586-018-0236-6.
(13) Mattedi, G.; Deflorian, F.; Mason, J. S.; de Graaf, C.; Gervasio, F. L. Understanding Ligand Binding Selectivity in a Prototypical GPCR Family. J. Chem. Inf. Model. 2019, 59, 2830–2836. https://doi.org/10.1021/acs.jcim.9b00298.
(14) van Westen, G. J. P.; van den Hoven, O. O.; van der Pijl, R.; Mulder-Krieger, T.; de Vries, H.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. J. Med. Chem. 2012, 55, 7010–7020. https://doi.org/10.1021/jm3003069.
(15) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. J. Cheminform. 2017, 9, 45. https://doi.org/10.1186/s13321-017-0232-0.
(16) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100-1107.
(17) Guo, D.; Mulder-Krieger, T.; IJzerman, A. P.; Heitman, L. H. Functional Efficacy of Adenosine A2A Receptor Agonists Is Positively Correlated to Their Receptor Residence Time. Br. J. Pharmacol. 2012, 166, 1846–1859. https://doi.org/10.1111/j.1476-5381.2012.01897.x.
(18) Cappellacci, L.; Franchetti, P.; Pasqualini, M.; Petrelli, R.; Vita, P.; Lavecchia, A.; Novellino, E.; Costa, B.; Martini, C.; Klotz, K.-N.; Grifantini, M. Synthesis, Biological Evaluation, and Molecular Modeling of Ribose-Modified Adenosine Analogues as Adenosine Receptor Agonists. J. Med. Chem. 2005, 48, 1550–1562. https://doi.org/10.1021/jm049408n.
(19) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. J. Comput. Aided. Mol. Des. 2015, 29, 885–896. https://doi.org/10.1007/s10822-015-9860-5.
(20) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. J. Mach. Learn. Res. 2008, 9, 2579–2605.
(21) Rucktooa, P.; Cheng, R. K. Y.; Segala, E.; Geng, T.; Errey, J. C.; Brown, G. A.; Cooke, R. M.; Marshall, F. H.; Doré,

A. S. Towards High Throughput GPCR Crystallography: In Meso Soaking of Adenosine A2A Receptor Crystals. Sci. Rep. 2018, 8, 41. https://doi.org/10.1038/s41598-017-18570-w.
(22) Kim, J.; Wess, J.; van Rhee, A. M.; Schöneberg, T.; Jacobson, K. A. Site-Directed Mutagenesis Identifies Residues Involved in Ligand Recognition in the Human A(2a) Adenosine Receptor. J. Biol. Chem. 1995, 270, 13987–13997.
(23) Sciabola, S.; Stanton, R. V; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S.; Xi, H. Predicting Kinase Selectivity Profiles Using Free-Wilson QSAR Analysis. J. Chem. Inf. Model. 2008, 48, 1851–1867. https://doi.org/10.1021/ci800138n.
(24) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. J. Med. Chem. 2012, 55, 5165–5173. https://doi.org/10.1021/jm300131x.
(25) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. J. Chem. Inf. Model. 2016, 56, 1936–1949. https://doi.org/10.1021/acs.jcim.6b00290.
(26) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Model. 2005, 45, 786–799. https://doi.org/10.1021/ci0500379.
(27) Zhao, L.; Xiang, Y.; Song, J.; Zhang, Z. A Novel Two-Step QSAR Modeling Work Flow to Predict Selectivity and Activity of HDAC Inhibitors. Bioorg. Med. Chem. Lett. 2013, 23, 929–933. https://doi.org/https://doi.org/10.1016/j.bmcl.2012.12.067.
(28) Frey, B. J.; Dueck, D. Clustering by Passing Messages Between Data Points. Science (80-.). 2007, 315, 972 LP – 976. https://doi.org/10.1126/science.1136800.
(29) Dassault Systèmes BIOVIA. BIOVIA Pipeline Pilot. Dassault Systèmes: San Diego 2018.
(30) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; WisWedel, B. KNIME: The Konstanz Information Miner. In Studies in Classification, Data Analysis, and Knowledge Organization; Springer, 2007.
(31) van Westen, G. J. P.; Gaulton, A.; Overington, J. P. Chemical, Target, and Bioactive Properties of Allosteric Modulation. PLoS Comput. Biol. 2014, 10, e1003559. https://doi.org/10.1371/journal.pcbi.1003559.
(32) Fawcett, T. An Introduction to ROC Analysis. Pattern Recognit. Lett. 2006, 27, 861–874. https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010.
(33) Schrödinger. Schrödinger Maestro Release 2017-4. Schrödinger: LLC, New York, NY 2017.

Chapter five

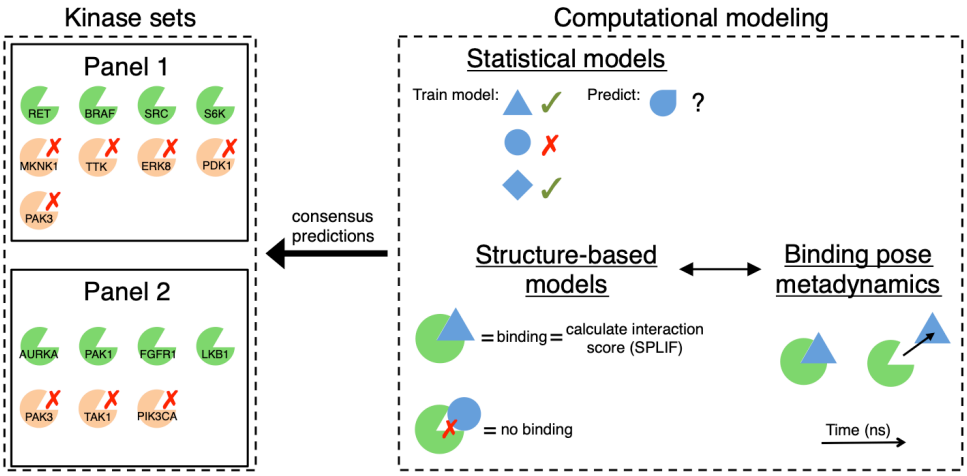
Successive Statistical and Structure-Based Modeling to Identify Chemically Novel Kinase Inhibitors



Lindsey Burggraaff, Eelke B. Lenselink, Willem Jespers, Jesper van Engelen, Brandon J. Bongers, Marina Gorostiola González, Rongfang Liu, Holger H. Hoos, Herman W.T. van Vlijmen, Adriaan P. IJzerman, and Gerard J.P. van Westen. Journal of Chemical Information and Modeling. 2020. doi: 10.1021/acs.jcim.9b01204.

Abstract

Kinases are frequently studied in the context of anti-cancer drugs. Their involvement in cell responses such as proliferation, differentiation, and apoptosis, makes them interesting subjects in multi-target drug design. In this study a workflow is presented that models the bioactivity spectra for two panels of kinases: 1) inhibition of RET, BRAF, SRC, and S6K, while avoiding inhibition of MKNK1, TTK, ERK8, PDK1, and PAK3; 2) inhibition of AURKA, PAK1, FGFR1, and LKB1, while avoiding inhibition of PAK3, TAK1, and PIK3CA. Both statistical and structure-based models were included, which were thoroughly benchmarked and optimized. A virtual screening was performed to test the workflow for one of the main targets, RET kinase. This resulted in 5 novel and chemically dissimilar RET inhibitors with remaining RET activity of < 60% (at a concentration of 10 μM) and similarities with known RET inhibitors from 0.18 to 0.29 (Tanimoto, ECFP6). The four more potent inhibitors were assessed in a concentration range, and proved to be modestly active with a pIC_{50} value of 5.1 for the most active compound. The experimental validation of inhibitors for RET strongly indicates that the multi-target workflow is able to detect novel inhibitors for kinases, and hence this workflow can potentially be applied in polypharmacology modeling. We conclude that this approach can identify new chemical matter for existing targets. Moreover, this workflow can easily be applied to other targets as well.



Introduction

Compound promiscuity can be leveraged to develop multi-target drugs. Such multi-target drugs can replace existing multi-drug treatments, while maintaining the therapeutic effect.[1] One of the advantages of a multi-target drug, or single-drug treatment is that no drug-drug interactions occur, making the treatment less risky and harmful for the patient.[2] However, since multi-target drugs are designed to bind to multiple proteins, they may tend to be more promiscuous as well. Therefore, when developing multi-target compounds, off-target binding and pathways should also be considered.

In the light of the "Multi-Targeting Drug" DREAM challenge,[3] bioactivities were computationally modeled for two panels of kinases. The first panel was based on treatment of medullary thyroid carcinoma, where kinases RET, BRAF, SRC, and S6K, should be inhibited and MKNK1, TTK, ERK8, PDK1, and PAK3, should not be affected.[4,5] RET was considered the main on-target kinase in this panel and thus was prioritized over other kinases in the panel. The second panel was based on tauopathies in neurogenerative disease: compounds should inhibit AURKA, PAK1, FGFR1, and LKB1, and not bind to PAK3, TAK1, and PIK3CA.[3] Since the main on-targets AURKA and PAK1, and additional on-targets FGFR1 and LKB1, are targeted in the central nervous system, compounds for panel 2 kinases should additionally be able to pass the blood-brain barrier.

Box 5.1 The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge is a contest organized by a community of researchers and is open to participants worldwide. Starting from 2006, every year multiple challenges are organized aimed at a different subject within biological science and human health. A common motive throughout the challenges is that it is promoted to find a solution using computational models and techniques. The "multi-targeting drug DREAM challenge"[3] was part of the 12th year edition of DREAM challenges.

This study describes a rigorous workflow to model the bioactivity spectra of compounds in kinases (**Figure 5.1**) and identify novel inhibitors. Every step in the workflow was extensively benchmarked, and each model was validated prior to virtual screening. A consensus scoring approach was used to rank virtual screening compounds, and only compounds with a Tanimoto similarity (ECFP6 (extended-connectivity bit string, diameter 6))[6] < 0.4 were considered to make sure that existing active molecules would not be 'rediscovered'. This consensus approach encompassed statistical modeling techniques such as quantitative structure-activity relationship (QSAR) models and proteochemometric (PCM) modeling.[7–9] Moreover, structure-based docking and pose metadynamics were applied.[10] Next to compound ranking, this approach was also used to exclude inactive-predicted compounds from virtual screening along the way. Fast machine learning models were applied to discard compounds early in the workflow. Subsequently, slower, but information rich, structure-based models were applied that consequently only had to score a smaller fraction of compounds compared to the entire initial virtual screening set. In this way, millions of compounds were screened rapidly and scored accurately.

In addition to the extensive benchmark validations, the workflow was also validated in vitro for one of the main on-target kinases. Forty-six compounds were experimentally validated for RET, of which 15 were selected based on the consensus approach,[6] based on statistical models only, and 25 compounds based on structure-based modeling. This resulted in a total of 5 inhibitors causing RET activity < 60% at a concentration of 10 μM. The four most potent compounds were further inspected and their IC_{50} were determined. Although the most potent RET inhibitor

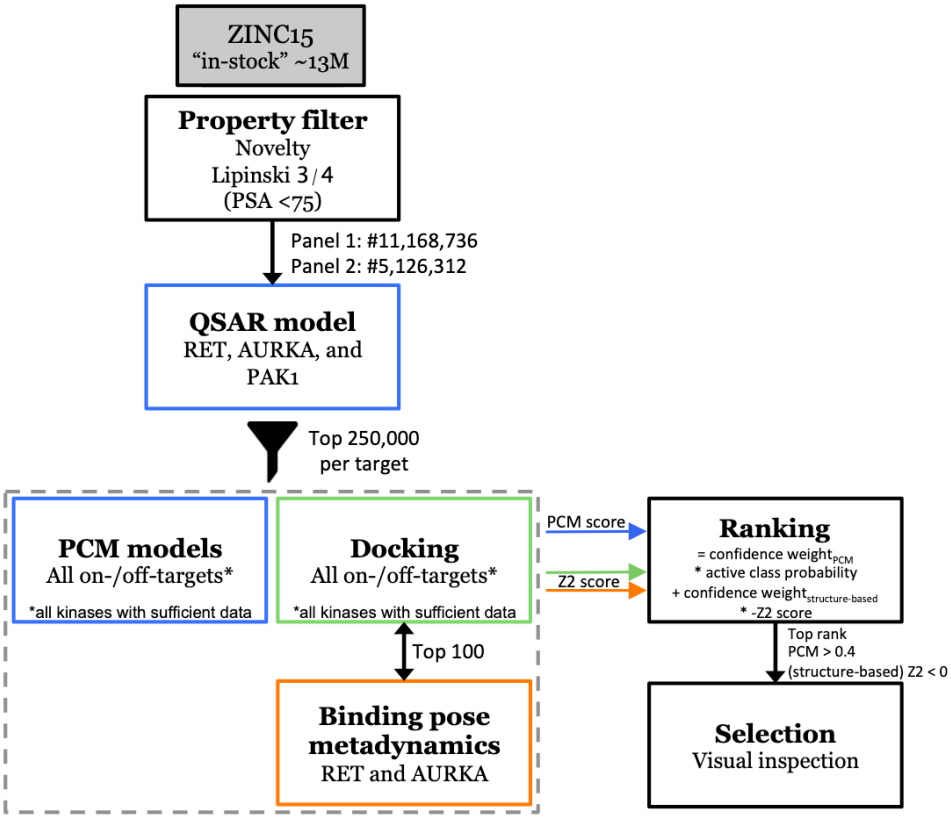


Figure 5.1. Virtual screening workflow for the two panels of kinases. Statistical models (blue), structure-based models (green), and molecular dynamics (orange) were applied to rank the virtual screening compounds.

was modestly active with a pIC_{50} of 5.1 ±0.1, all compounds were chemically distinct from known RET inhibitors with Tanimoto (ECFP6) similarities smaller than 0.29. Therefore, the identified inhibitors in this study provide a new starting point for hit/lead optimization based on novel scaffolds.

Results

Data curation and filtering

Compound bioactivity data was derived from different sources to increase data availability for the benchmark sets in model training and validation. Kinase compound data was retrieved from the following sources: ChEMBL database (version 23)[11], publicly available sets from Eidogen[12], and ExCAPE-DB[13]. The data was curated by standardizing chemical structures and bioactivity values (see Methods section for details). In this way a dataset was constituted that contained compound bioactivity information for 512 kinases.[i] This dataset was used for training and testing of statistical models. Since validation was performed using cross-validation, the dataset was divided into subsets; five subsets were created based on chemical clustering of the compounds per target (see Methods section for details). The biggest subset was used in training of every

i dataset available at doi: 10.4121/uuid:6af1d9de-281f-4221-b7e1-e7c01b90dfe0.

model, whereas the remaining four subsets were used rotationally for testing. The subsets that were not used in testing in the specific iteration were added to the training set.

Additionally, separate active/inactive/decoy datasets were constructed for validation of structure-based models. These benchmark sets were constructed for each panel kinase separately. Data for these benchmark sets were derived from ChEMBL (version 23), which was curated and filtered to construct a benchmark set for each kinase containing inactive and chemically diverse active compounds: only compounds with measured pChEMBL[14] values were included in the benchmark sets and a restrain was set for the number of active compounds. Up to a maximum of 100 actives per kinase were included, whereas the number of inactive compounds was not limited. The 100 active compounds were selected based on chemical diversity by clustering the compounds with pChEMBL value > 6.5 into 100 clusters and selecting only the cluster centers. Additionally, DUD-E[15] decoys were added to each benchmark set. These decoys have similar physicochemical properties as active ligands, but differ in chemical structure. The structure-based benchmark sets were smaller than the training and test sets of statistical models, but big enough to allow for validation of the models. The smaller size of the structure-based benchmark sets allowed for quicker model evaluation, resulting in validation of many protein structures.

The virtual screening set that was screened using both statistical- and structure-based models was derived from the ZINC15[16] database. All "in stock" compounds were filtered on drug-like properties by discarding compounds that did not adhere to 3 of 4 Lipinski rules.[17] Furthermore, compounds were filtered on novelty: compounds with Tanimoto similarity (ECFP6) > 0.4 compared to existing actives (pChEMBL > 5) on the kinases in the respective panel were excluded from the virtual screening set. The virtual screening set was additionally filtered for panel 2 kinases by including the likelihood of compound passing the central nervous system by only keeping compounds with polar surface area < 75 $Å^2$. This resulted in a virtual screening set for panel 1 of 11,168,736 compounds, and for panel 2 of 5,126,312 compounds.

Statistical models

Separate quantitative structure-activity relationship (QSAR) models were constructed for the main on-targets RET, AURKA, and PAK1, as a first filter for bioactive compounds. The models were validated with 4-fold cross-validation using standardized benchmark sets (see Methods for details). The benchmark sets were constructed per target and were extracted from the main statistical benchmark set containing 512 kinases. Chemical descriptors were calculated for every compound: FCFP4 (feature-connectivity bit string, diameter 4) fingerprints and physicochemical descriptors (**listed in table S5.1**). These descriptors describe the compounds and were used in training the models. The RET, AURKA, and PAK1 QSAR models were predictive with ROC (receiver operating characteristic) scores higher than random (**Table 1**). The ROC of the QSAR models was comparable between targets (ROC 0.76 ±0.01), whereas the Matthews correlation coefficient (MCC) varied slightly more (MCC 0.30 ±0.08).

The performances of the QSAR models were sufficient as a first filter for bioactive compounds, discarding the least active compounds and steering clear of the decision boundary. Virtual screening set 1 was screened using the RET QSAR, and virtual screening set 2 was screened using the AURKA and PAK1 QSAR models separately. Using the active class probability score, the most promising compounds (250,000 compounds per RET/AURKA/PAK1) were selected

to be further processed in the structure-based approaches. This pre-screening of compounds with a simple, but fast model decreased the number of compounds effectively. As a result, subsequent screening and scoring steps proceeded quicker, since fewer compounds had to be screened. The subsequent steps, proteochemometrics (PCM)[9] and structure-based modeling, were carried out in parallel. Additionally, QSAR models were constructed for the remaining kinases simultaneously. These QSAR models were compared to the more advanced PCM models to select the best approach for scoring. The performances of these QSAR models are included in **Table 1**.

The PCM models that were created were applied solely for the purpose of scoring and not as a filtering step. PIK3CA was excluded from modeling, as insufficient data was available to build and validate the models. The PCM models were constructed for each kinase separately and were based on the particular kinase and its L4 level family members as annotated in ChEMBL.[18] In addition to the compound descriptors used in QSAR modeling (FCFP4 fingerprints and physicochemical properties), the PCM models included protein descriptors that were based on a full kinome sequence alignment (see Methods for details). Initial PCM models were trained using default random forest settings: 300 trees, log2(m) features at every node in the tree, no maximum tree depth, a minimum of 2 samples to consider a node for splitting, and bootstrap enabled. This resulted in an overall performance of: MCC 0.22 and ROC 0.69 (average over all panel kinases). Since the PCM models were intended for scoring of compounds, the models were optimized to enhance predictive performance. Optimal settings were explored by tuning hyperparameters with random search, a basic approach to automated machine learning.[19] Approximately 500 random forest models were trained during optimization, resulting in the best model with performance MCC 0.25 and ROC 0.74, and following settings: 300 trees (fixed) with 43 features at every node in the tree, a maximum tree depth of 99, a minimum of 12 samples to consider a node for splitting, and bootstrapping disabled. The performance of the QSAR and PCM models per kinase are shown in **Table 5.1**.

Predictions for most kinases were comparable between QSAR and PCM modeling. However, for target PAK3, PCM clearly outperformed QSAR with MCC difference 0.18 and ROC difference 0.20. MCC could not be calculated for ERK8 due to a small dataset and consequently a lack of a predicted class (total number of compounds for ERK8 is 332). Using QSAR, the MCC displays a negative correlation for ERK8 (-0.12), which is also reflected by the ROC score that is worse than random (ROC < 0.5). PDK1 has high early enrichment (BEDROC (Boltzmann-Enhanced Discrimination of the Receiver Operating Characteristic)) with both QSAR and PCM. Although overall enrichment (ROC) for PDK1 is lower than the early enrichments, it is still good with ROC 0.71 (QSAR) and 0.72 (PCM). The average over all targets shows that PCM predicts slightly better than QSAR, with a similar MCC score of 0.25, but higher (BED)ROC scores: differences BEDROC = 0.06, and ROC = 0.06. Moreover, the nature of PCM models allows for extrapolation of bioactivities from related kinases to the target of interest. Therefore, it was hypothesized that the PCM models will perform better than QSAR models when applied to a more diverse chemical dataset such as the virtual screening sets. The performances of the PCM models of the main on-targets AURKA and PAK1 were higher than the average performance over all targets. However, main on-target RET had a PCM MCC value (0.15) that was lower than average (0.25). Nevertheless, (BED)ROC scores were higher than the average: RET BEDROC 0.64, RET ROC 0.76, average BEDROC 0.58, and average ROC 0.74. Moreover, all RET scores

Table 5.1. PCM performance per target (mean over 4-fold cross-validation).

	PCM			QSAR			Number of compounds	
Target	MCC	BEDROC (α = 20)	ROC	MCC	BEDROC (α = 20)	ROC	Active (pChEM-BL ≥ 6.5)	Inactive (pChEM-BL < 6.5)
RET	0.15	0.64	0.76	0.23	0.63	0.75	1492	416
BRAF	0.18	0.74	0.56	0.20	0.75	0.54	1119	1359
SRC	0.28	0.47	0.72	0.26	0.47	0.72	4642	2238
S6K	0.38	0.85	0.79	0.45	0.85	0.78	1662	685
MKNK1	0.09*	0.42	0.61	0.01	0.32	0.50	549	51
TTK1	0.22	0.45	0.78	0.26	0.44	0.75	663	276
ERK8	****	0.05	0.48	-0.12	0.02	0.35	302	30
PDK1	0.27*	0.85	0.72	0.31	0.86	0.71	579	536
PAK3	0.25	0.72	0.91	0.07	0.27	0.71	1204	53
AURKA	0.37	0.65	0.78	0.38	0.47	0.77	3165	1674
PAK1	0.32	0.74	0.86	0.28	0.66	0.77	712	114
FGFR1	0.41	0.70	0.85	0.77	0.71	0.82	2477	928
LKB1	0.57**	0.53	0.76	0.26*	0.45	0.63	429	47
TAK1	0.15***	0.27	0.68	0.12*	0.33	0.69	1204	53
Average	0.25	0.58	0.74	0.25	0.52	0.68	295	56

Asterisks indicate that no value could be determined due to lack of predicted (positive/negative) classes: * = 1 cross-validation failed, ** = 2 cross-validations failed, *** = 3 cross-validations failed, and **** = 4 cross-validations failed. Indicated in bold in each column is the best performing model for that given parameter.

were better than random indicating the predictive power of the model.
The settings that corresponded to the best performing model in 4-fold cross-validation were applied to train PCM models on the full dataset per kinase (including test set in training). The models were applied to score all the virtual screening compounds that passed the filtering step using QSAR models.

Structure-based models

Structure-based scoring was performed in addition to scoring of compounds using the PCM models. For many kinases,multiple crystal structures have been deposited in the PDB, and it is often not obvious which crystal structure should be used prospectively. Therefore we performed a rigorous benchmark to determine the best enriching crystal structure. The number of validated crystal structures for all targets ranged from 1 to 135, excluding ERK8 and PAK3 for which no crystal structures were available at the time (December 2017). The crystal structures that were deemed suitable for virtual screening were X-ray protein structures that contained a co-crystalized orthosteric ligand. A total of 499 crystal structures were benchmarked using corresponding compound benchmark sets that were composed for each target separately. The benchmark sets containing active compounds, inactive compounds, and decoys, were docked into the orthosteric binding pockets, from which a docking score was derived for each compound. The decoys were considered as inactive compounds when calculating actives enrichment for each crystal structure. Enrichment was calculated based on best docking score per compound. It was observed that actives enrichment varied greatly between different structures of the same kinase: ROC ranging

from 0.58 (PDB: 2IVS) to 0.77 (2IVU) for RET, 0.47 (5OSF) to 0.80 (2XNG and 4O0W) for AURKA, and 0.65 (3Q4Z and 4O0T) to 0.95 (5IME) for PAK1. The models per target were ranked based on the sum of the overall (ROC) and early enrichment (BEDROC, α = 160.9). Models with a score ROC+BEDROC ≥ 1 (e.g. ROC 0.40 + BEDROC 0.60) were considered sufficient for prediction of active compounds.

Kinases LKB1 and MKNK1 only had a few crystal structures available (1 and 2 structures, respectively) of which performance was low (ROC+BEDROC: 0.50 and 0.72 respectively). Moreover, kinase S6K of which 18 structures were available, only reached maximum performance of ROC+BEDROC 0.80. Therefore, additional protein structures were created for these kinases by application of induced-fit docking. In contrast to docking, induced-fit docking accommodates the ligand and additionally re-orientates the side chains of binding pocket residues. This allows the residues to change conformation in order for the ligand to fit into the binding pocket. Five ligands for each LKB1, MKNK1, and S6K, were docked with induced-fit docking into the respective binding pocket. Subsequently, the ligands were removed from the protein, keeping only the altered protein structures. This resulted in 95 additional protein structures for LKB1, 44 for MKNK1, and 74 for S6K. These additional structures were validated using the same benchmark sets as used for the initial structures. The protein structures that were created using induced-fit docking varied in performance from ROC+BEDROC 0.27 to 1.07. For all three kinases, an induced-fit structure was generated that outperformed the initial crystal structure. The best performance for LKB1 was ROC+BEDROC 0.99, for MKNK1 it was 1.07, and for S6K it was 0.97.[i]

Based on the docking performances of all models, five protein structures per target were selected for structural protein-ligand interaction fingerprints (SPLIF)[20] calculations. With SPLIF, interactions between the binding pocket residues and (docked) ligand are calculated, resulting in a SPLIF score that indicates the similarity between the interactions of the (docked) ligand compared to a reference compound. In this case, the reference compound consists of the co-crystalized ligand in the corresponding structure. The structures for SPLIF calculations were selected based on diversity of the co-crystalized ligands per kinase. The diversity of these ligands was assessed by clustering the ligands using affinity propagation clustering.[21] The cluster centers of the structures with highest ROC+BEDROC in docking were selected as reference. Consequently, the corresponding protein structures and docked benchmark set poses were thus used in SPLIF calculations. The SPLIF similarity scores for each benchmark compound were used to calculate actives enrichment based on SPLIFs. Especially for MKNK1 actives enrichment increased significantly compared to enrichment based on docking scores: ROC+BEDROC SPLIF 1.57 versus ROC+BEDROC docking 1.07. Based on early enrichment, BEDROC (α = 160.9), the performance using SPLIF scores increased for 9 of 13 kinases (**Figure 5.2**). However, for targets BRAF, PDK1, PAK1, and FGFR1 docking scores enriched (BEDROC (α = 160.9)) better than SPLIF scores.

An ensemble score was constituted, which combined docking and SPLIF scores. This ensemble score, the Z2 score,[22] was based on only docking scores, only SPLIF scores, or a combination of docking and SPLIF scores (see Methods for details). Prior to calculation of Z2-scores, the docking and SPLIF scores were normalized to Z-scores (more negative Z-score means better

i The protein structures resulting from induced-fit docking are available at doi: 10.4121/uuid:9e61b6a6-88e5-4a18-ba19-dbf1bbdd656a.

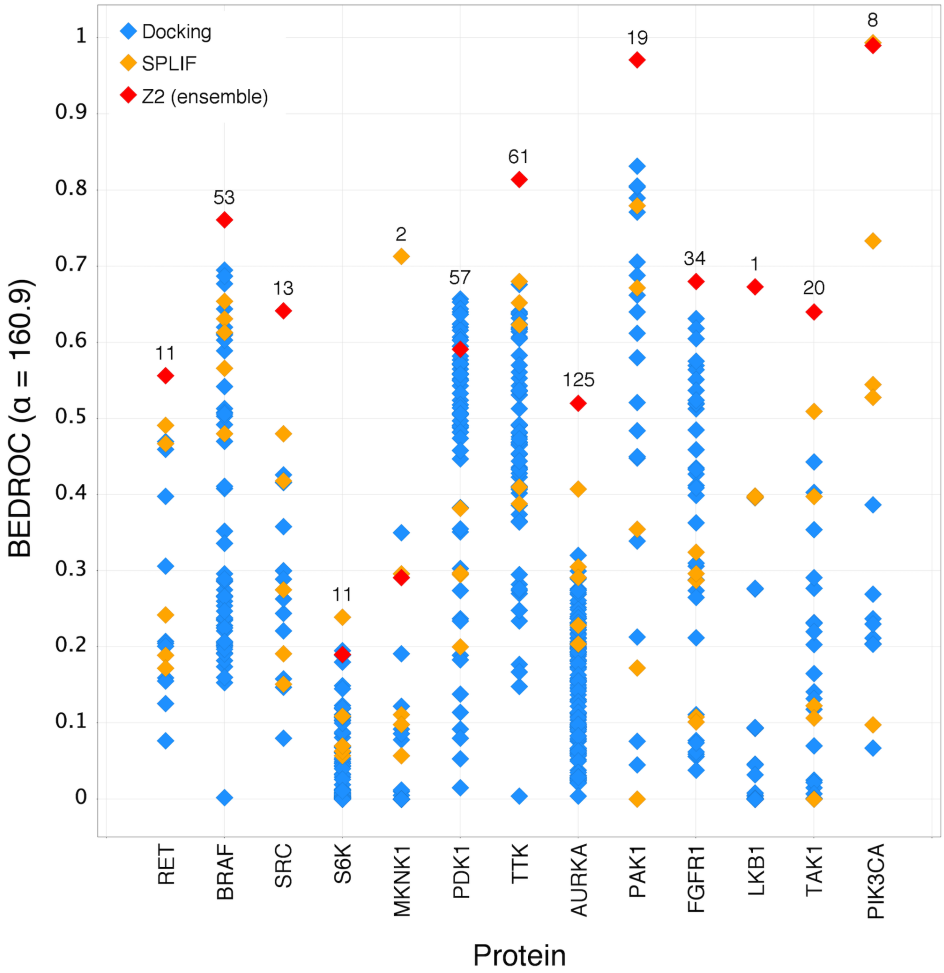


Figure 5.2. Early enrichment of actives, BEDROC (α = 160.9), per crystal structure for each target. Enrichment reached by docking scores (blue), SPLIF scores (orange), and Z2-scores (red), are shown for all 13 kinases that had a crystal structure available. Numbers on top indicate the number of crystal structures for each kinase.

binding). This normalization was done with respect to the actives from the benchmark set for a given kinase. The early enrichments of active compounds based on Z2-scoring increased performance for 10 of the 13 targets (**Figure 5.2**). The best (ensemble of) models per target are listed in **Table S5.2**. The best performances of the main on-targets RET, AURKA, and PAK1 were reached by Z2-scoring. The performance of PAK1 was very good with BEDROC (α = 160.9) 0.97. However, it should be noted that the number of compounds in this benchmark set was low due to lack of data (63 active compounds, 10 inactive compounds and 3886 decoys). Therefore, this model may not be representative when applied to a virtual screening set. The best BEDROC performances of RET and AURKA were 0.56 and 0.52 respectively. Furthermore, the overall performance for these targets were good with ROC 0.85 for RET and ROC 0.82 for AURKA.

Prospective structure-based docking

The virtual screening compounds resulting from QSAR filtering were docked into the protein structures that resulted in the best performances in benchmarking. Not the entire virtual screening set could be docked on all kinases because of time constraints. Therefore the top 250,000 compounds for RET, AURKA, and PAK1, with highest active-class probabilities, were selected for docking. This included all active-predicted compounds for RET and AURKA (167,828 for RET and 315 for AURKA), and additional, consecutively ranked compounds that did not reach the QSAR activity threshold (active class probability > 0.5). For PAK1 the top 250,000 compounds were selected from a total of 298,163 active-predicted compounds. Subsequently, the compounds were assigned a structure-based Z2-score, with the exception of S6K, MKNK1, and PDK1 for which SPLIF or docking score gave the best BEDROC ($\alpha = 160.9$) performance and thus Z2-scores were replaced with the respective score. The best scoring method for S6K underperformed with a BEDROC of 0.24. Although this model was not discarded immediately, the poor performance was taken into account later when reliability weights were assigned to the corresponding method and kinase.

Binding pose metadynamics

Binding pose metadynamics[23,24] was performed to re-score the docked poses and scores of compounds. Since binding pose metadynamics is a time-consuming modeling technique, only the main on-targets (RET, AURKA, and PAK1) were subjected to this method. Binding pose metadynamics measures the persistence of ligand-protein interactions and the movements of the ligand's heavy atoms, which are sampled during variation of the complex's free energy states throughout the simulation. The result of binding pose metadynamics is a metadynamics-composition score. This composition score was calculated for the top 100 compounds (based on docking score) from the benchmark set of each included kinase. The docking scores from which these top 100 compounds were chosen were derived from a single protein structure per target to allow for easy and direct comparison. The selected protein structures had the best actives enrichment based on docking and included 2IVU for RET, 2BMC for AURKA, and 4EQC for PAK1. The composition score was added to the docking score, resulting in a combined score. The actives enrichment for the targets was re-evaluated using this new score. It was observed that performance (based on the top 100 compounds) of PAK1 did not increase (ROC+BEDROC difference 0.01). However, actives enrichment for RET and AURKA increased with performances (ROC+BEDROC) 1.27 for RET and 1.76 for AURKA, compared to initial (docking) performance of 1.26 for RET and 0.77 for AURKA. Therefore, for RET and AURKA the docking scores of virtual screening compounds were re-scored with metadynamics-composition scores, resulting in an indirect re-ranking of the top 100 virtual screening compounds, which was based on Z2-scores.

Polypharmacology ranking of compounds

The virtual screening compounds were scored with both a statistical model score (PCM score) and structure-based score. A final compound rank per target was constituted by adding a weight to the predictions made by statistical models and structure-based models (**Table 5.2**). The ranking order of compounds for off-targets, i.e. kinases compounds should not interact with, was reversed: ranking compounds with high predicted activity as lowest, and compounds with worst predicted activity as highest. The weights of the statistical PCM model were based on the

ROC score corrected for the size of the training dataset per target. The structure-based models were considered to be better suited to select novel chemistry and were therefore assigned higher weights than the PCM model weights: structure-based models were attributed more weight compared to equally performing PCM models (same ROC). The weights of the structure-based models were calculated by taking the sum of BEDROC+ROC. These weights were reduced by penalizing the models when induced-fit structures were used and when the numbers of compounds in the benchmark sets were insufficient (for details see Methods). The weights were subsequently used to rank the compounds per target: Structure-based Z2-scores were multiplied by the structure-based weight, PCM predicted class probability was multiplied by the statistical model weight, and as final step the derived scores were summed up to retrieve a final rank per target.

Based on overall target weights (**Table 5.2**), predictions for kinases LKB1, ERK8, PAK3, TAK1, and PIK3CA, were not very reliable. However, predictions for the more important main on-targets RET, AURKA, and PAK1 were considered to be reliable. The workflow was evaluated by the selection and experimental validation of compounds for one of the main on-targets: RET. The consensus approach was used to select 15 highly ranked virtual screening compounds for RET (all compounds had PCM score > 0.4 and structure-based score Z2 < 0), which were validated in vitro. Moreover, the performance of the different approaches was compared on the RET models by additionally selecting compounds based on the predictions of only statistical models, and only structure-based models.

Table 5.2. Weights assigned to each target per modeling technique.

Model	**On-targets**				**off-targets**				
Panel 1 kinases	**RET**	**BRAF**	**SRC**	**S6K**	**MKNK1**	**TTK**	**ERK8**	**PDK1**	**PAK3**
PCM	0.54	0.14	0.90	0.67	0.13	0.41	-0.02	0.35	0.70
Structure Based	1.44	1.66	1.56	0.48	1.42	1.79	n.a.	1.45	n.a.
Total	1.98	1.80	2.46	1.15	1.55	2.20	-0.02	1.80	0.70
Panel 2 kinases	**AURKA**	**PAK1**	**FGFR1**	**LKB1**	**TAK1**	**PIK3CA**	**PAK3**		
PCM	0.93	0.50	0.98	0.28	0.16	n.a.	0.70		
Structure Based	1.34	1.23	1.62	0.04	0.46	0.97	n.a.		
Total	2.27	1.73	2.60	0.32	0.62	0.97	0.70		

Experimental validation

The models were evaluated by experimental validation of predicted actives for RET kinase. A total selection of 46 compounds was purchased and validated for RET inhibition. These compounds were selected by different criteria: 6 compounds based on predicted activity by QSAR and PCM modeling, 25 compounds based on structure-based docking (including re-scoring by binding pose metadynamics) and SPLIF scoring, and 15 compounds based on consensus scoring of statistical and structure-based models.

Compounds that were selected using only statistical models satisfied the set active-class threshold criteria of both QSAR (active probability > 0.6) and PCM (active probability > 0.5). The structure-based thresholds were docking score < -8 and SPLIF score > 0.25. Additionally, structure-based compounds with a docking score < -10 were selected that did not necessarily adhere to the SPLIF score criteria. Furthermore, compounds that were selected based on consensus scoring fitted the following thresholds: statistical predictions PCM > 0.4, and structure-based score Z2 < 0. The compounds that were selected based on structure-based predictions or with consensus scoring were additionally inspected visually by checking the 3D docking pose and interaction with the hinge region.

The 46 compounds were first tested using single point measurements at two concentrations: 10 μM and 0.1 μM (**Table S5.3**). The top ten compounds showing RET inhibition (RET activity < 80%, concentration 10 μM), were based on either consensus scoring or structure-based scoring. None of these compounds was from statistical scoring only. Five compounds showed inhibitory activity (RET activity < 60%) for RET at a concentration of 10 μM, of which one compound also showed slight RET inhibition at concentration 0.1 μM (RET activity < 100%). The activities of the four most potent hits were assessed more accurately by a potency determination in triplicate, yielding pIC_{50} values. The inhibitors were modestly active. ZINC33008650 was the best inhibitor with a pIC_{50} of 5.1 ±0.1, followed by ZINC72312837 (pIC_{50} 4.6 ±0.2), ZINC12324934 (pIC_{50}

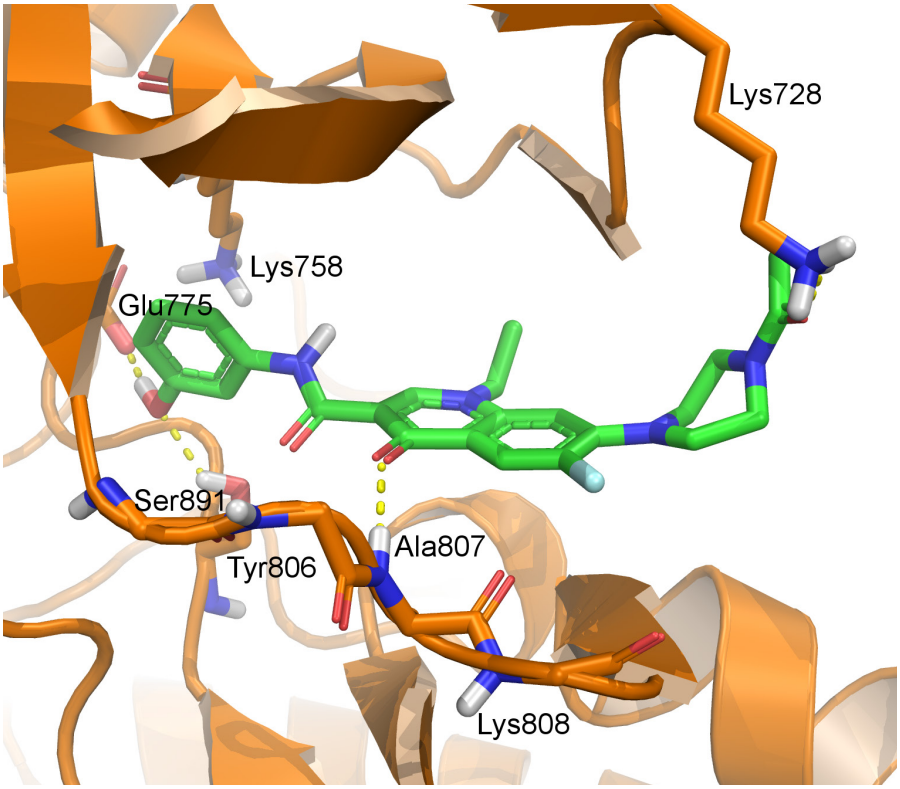


Figure 5.3. ZINC12324934 (green) docked into the orthosteric binding pocket of RET (orange) (PDB:2IVU). Hydrogen bonds are displayed as yellow dotted lines.

4.6 ±0.2), and ZINC9518200 (pIC_{50} 4.0 ±0.2). The docked pose of ZINC12324934 in RET suggests that the compound is an orthosteric binder with potential to bind to the hinge region, as evidenced by a hydrogen bond interaction with the backbone of Ala807 (**Figure 5.3**). Additional hydrogen bonds are observed between the ligand and Glu775, Ser891, and Lys728. Although the inhibitors showed modest activity, their chemical diversity compared to known RET inhibitors was high since the compounds were pre-filtered on novelty by selecting on Tanimoto similarity < 0.4 (compared to inhibitors with pChEMBL value ≥ 5).

Bioactivity spectra prediction

The bioactivity profile of the five most active inhibitors for RET was predicted for all kinases in the panel. The bioactivity spectra for both on- and off-targets are shown in **Table 5.3**, where positive values indicate binding and values below and including zero indicate no binding. Based on the entire predicted bioactivity spectra, the most potent RET inhibitors comply better with inactivity on off-targets than activity on on-targets. However, the predictions are not equally reliable for each kinase (based on the previously assessed weights shown in **Table 5.2**). Nevertheless, assuming that the predictions of the kinases that scored equally well or better compared to RET (weight 1.98) are the most accurate, conclusions can be drawn for on-target SRC and off-target TTK (weights 2.46 and 2.20, respectively). Although the RET inhibitors were not predicted active for SRC, inactivity was predicted for off-target TTK. The compounds were selected based on bioactivity for RET, and as a result may not show optimal bioactivity spectra. However, the novelty filter to which the compounds were subjected prior to virtual screening included all of the kinases in the panel (as opposed to only RET). Therefore, all predicted interactions might indicate novel starting points for future research.

Table 5.3. Predicted bioactivity spectra for the panel 1 kinases of the five most potent RET inhibitors.

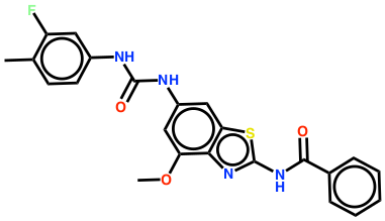
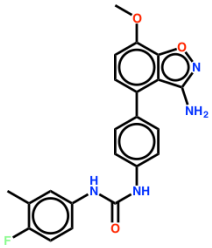
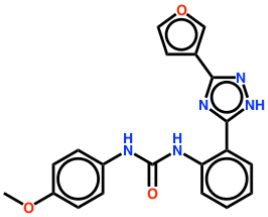
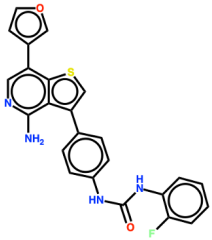
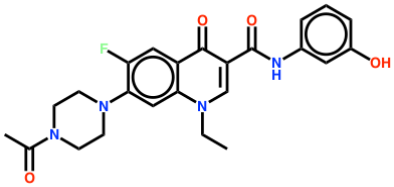
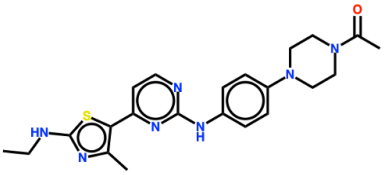
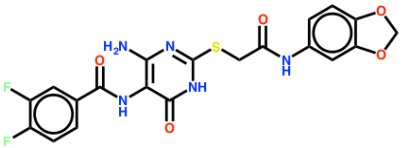
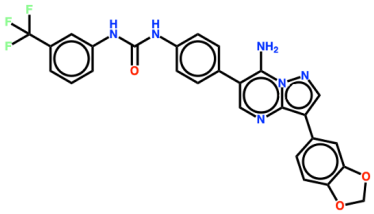
Compound	On-targets				Off-targets				
	RET	BRAF	SRC	S6K	MKNK1	TTK	ERK8*	PDK1	PAK3*
ZINC33008650	1	0	0	0	-2	-2	0	-2	0
ZINC12324934[§]	0	0	0	0	0	0	0	0	0
ZINC9518200	2	0	0	0	-1	-1	0	-1	0
ZINC72312837	2	0	0	0	-1	-1	0	-2	0
ZINC65184824	1	-1	0	0	-2	-1	0	-2	0

*Predictions based on limited structure-based data (no crystal structure available). [§]Compound was selected on RET docking score only, therefore no Z2-score was available for this compound and no structure-based weight could be assigned.

Manual inspection of hits

The four most potent hit compounds resulting from virtual screening and experimental validation were inspected based on their novelty and patentability. Novelty was re-evaluated by similarity searching (Tanimoto ECFP6) on a more strict threshold for RET compounds (pChEMBL ≥ 4) in the most recent version of ChEMBL (version 25) and patentability was checked in SureChEMBL[25] (similarity > 0.9). The most similar RET actives were: CHEMBL1979093 for ZINC33008650 (similarity 0.24, RET pChEMBL 7.2), CHEMBL1983715 for ZINC12324934 (similarity 0.18, RET pChEMBL 6.9), CHEMBL1977134 for ZINC9518200 (similarity 0.20,

Table 5.4. Novel RET inhibitors and their most similar RET actives in ChEMBL based on Tanimoto (ECFP6).

Compound	Closest similar (pChEMBL ≥ 4)	Similarity (Tanimoto ECFP6)
 ZINC33008650 pIC_{50} 5.1 ±0.1	 CHEMBL1979093 RET pChEMBL 7.2	0.24
 ZINC72312837 pIC_{50} 4.6 ±0.2	 CHEMBL1965570 RET pChEMBL 7.6	0.29
 ZINC12324934 pIC_{50} 4.6 ±0.2	 CHEMBL1983715 RET pChEMBL 6.9	0.18
 ZINC9518200 pIC_{50} 4.0 ±0.2	 CHEMBL1977134 RET pChEMBL 8.4	0.20

RET pChEMBL 8.4), and CHEMBL1965570 for ZINC72312837 (similarity 0.29, RET pChEMBL 7.6) (**Table 5.4**). None of the compounds was listed as patented in SureChEMBL.

Additionally, novelty of the chemical scaffolds was checked, which were compared based on Bemis-Murcko[26] scaffold trees with known active compounds for RET in ChEMBL. None of the four hit compounds was clustered in the same scaffold cluster as a known active, supporting the novelty of the inhibitors. Although the hit compounds were identified as modestly active binders, further SAR around these inhibitors may yield more potent derivatives. The novel scaffolds may be explored along different vectors to enhance affinity of the ligands and may reveal a new chemical space for RET inhibitors.

Discussion

An elaborate virtual screening process was constituted in which both statistical and structure-based models were applied that were all carefully tuned and validated. Statistical models were validated using as many compound data as possible, with attention paid to chemical diversity of the subsets in cross-validation. Structure-based models were slower than the statistical models and therefore smaller benchmark sets were used for validation of these models. These benchmark sets were composed of diverse actives to cover a large chemical space, irrespective of the smaller size of the dataset. By application of statistical QSAR models and structure-based docking successively, computational time was used efficiently in virtual screening: compounds that had the least predicted activity probability for the main on-targets based on the QSAR models were excluded from docking. The consensus scoring approach, in which statistical PCM scores were combined with structure-based predictions, resulted in sets of active-predicted compounds per kinase. However, compounds were excluded from bioactivity modeling if the respective kinase had a poor reliability weight. This reliability weight was based on the performances per model per target derived from the benchmarking steps.

It was observed that the lack of compound data resulted in lower performance of statistical models in particular, as seen for kinase ERK8. Although structure-based models also performed worse for targets with less compound data and less crystal structures, this could partially be resolved by creation of additional protein structures using induced-fit docking and alternative scoring with SPLIF and Z2. However, alternate structures may also have been created using more elaborate ligand-protein sampling methods such as molecular dynamics.[27] Interestingly, performance between different structures of the same kinase varied greatly, rendering the benchmarking of multiple protein structures necessary.

Based on the final reliability scores, kinases RET, BRAF, SRC, MKNK1, TTK, PDK1, AURKA, PAK1, and FGFR1, are best used in bioactivity modeling, whereas predictions for kinases S6K, ERK8, PAK3, LKB1, TAK1, and PIK3CA, may not be accurate enough and thus should be excluded for now. Although the models of the first kinases are considered reliable, identification of compounds that fit the desired bioactivity profiles on these targets remains challenging. Since hit rates are rarely 100%, compounds with a well-predicted profile would need experimental validation. The hit rate for the experimental validation for RET in this study was 11% over 46 compounds, using $pIC_{50} \geq 4$ (or RET activity < 60% at a concentration of 10 μM) as a threshold for active compounds. On the hypothesis that similar hit rates will be achieved for the other

kinases, the probability of a compound being active on a number of kinases thus decreases with every additional target. Therefore, the number of experimentally validated compounds should be expanded, to increase the chance of finding a compound that fully fits the desired bioactivity profile.

The five hits that proved active on RET kinase were found with either consensus scoring or structure-based modeling. Although the number of compounds validated using statistical models is smaller, it is plausible that statistical modeling by itself is not a strong enough predictor to identify really novel compounds.[28] Since the virtual screening compounds were pre-filtered on novelty (Tanimoto ECFP6 < 0.4) with known actives, the models were challenged with identification of chemical dissimilar actives.[29] This is a difficult task for statistical models as they rely on chemical patterns of known actives, and therefore dissimilar compounds may lie outside the applicability domain of these models.[29] To resolve this issue, the chemical space can be expanded by addition of chemical diversity.[8] Although this requires biological experiments to validate (in)activity of new compounds, an iterative screening approach may be applied to expand the chemical space efficiently and cost-effectively.[30] Nevertheless, identification of novel chemistry without the need for addition of much experimental data may be achieved through structure-based modeling, as these models are not particularly biased towards known chemistry. However, bias in structure-based models may indirectly be present, as the best performing protein structures were based on enrichment of known actives. Nevertheless, this is a minor issue compared to the limited scope and bias of chemical space of statistical models.

The applied workflow in this study employs a filtering step through which compounds are discarded that were categorized as inactive by statistical QSAR models. Based on the previous statement on applicability domain, it is assumed good inhibitors can be wrongly categorized as "inactive" and neglected by the statistical model. One might reason that docking of all compounds may be a more effective method to identify novel inhibitors, as docking of 170 million compounds in proteins AmpC β-lactamase and the D4 dopamine receptor resulted in novel chemical scaffolds.[31] However, docking of millions of compounds on multiple proteins is a very time consuming task (~170,000 compounds per day for 1 RET crystal structure (24 cores)), which is unfeasible without the required computational resources. Thus, considering the scope of the task, the possibility that the QSAR models discard 'good' compounds is accepted, as the QSAR models decrease the runtime of the workflow and make the task comprehensive.
As a final remark, it should be mentioned that the workflow applied to kinases did not implement orthosteric or allosteric binding of compounds. Therefore, inhibitors were not tuned to bind to the DFG-in or DFG-out kinase conformation, a conformational change that influences inhibitor binding greatly.[32] Although structure-based docking was focused on the ATP-binding pocket, the most optimal crystal structures were selected based on benchmark sets that were not filtered for DFG-out and allosteric binders. As a consequence, crystal structures may have been selected that enriched DFG-out binders better than DFG-in binders. Moreover, since the statistical models were trained on sets that were not filtered for DFG-out binders, these models were also not able to distinguish DFG-in from DFG-out binders. As a result, it is undetermined whether the five hits from the screening workflow bind to the DFG-in conformation of RET. The docked poses of the hits do not constitute optimal hinge-binding, suggesting that it is plausible that they may be DFG-out binders. Moreover, two of the five most potent hit compounds contained a urea-motif, a motif that is often associated with DFG-out binders.[32,33] To capture the binding type of

compounds, machine learning models could have been used to predict the type of compound as an additional score, something that will be considered for future work.[34,35]

Conclusion

An extensive workflow was designed to predict compound activity in kinases and to model the compounds' bioactivity spectra in kinases. The workflow can easily be expanded to other targets as well. By combining statistical and structure-based modeling, processing speed was optimized, while the accuracy of predictions was preserved. Every single kinase target was validated separately, which enabled reliability weights to be assigned to the predictions for every target. The workflow was experimentally validated by testing predictions made for RET kinase, a target with a good reliability score. A selection of 46 compounds was tested in vitro, of which 5 compounds showed RET inhibition (activity < 60%) at a concentration of 10 μM. The four most potent inhibitors had pIC_{50} values ranging from 4.0 to 5.1. The Tanimoto similarities (ECFP6) of these inhibitors with known RET actives was ≤ 0.29. Moreover, the compounds contained unique chemical scaffolds, underscoring the true novelty of these inhibitors.

Methods

Dataset statistical models

Training and validation sets for statistical models were constituted from compound information derived from ChEMBL (version 23)[11], publicly available sets from Eidogen[12], and ExCAPE-DB[13]. The compounds with experimental bioactivity for any kinase were filtered on molecular weight < 700, duplicates were removed, and compounds were standardized using BIOVIA Pipeline Pilot 2016[36]: salts were removed, largest fragment was kept, stereochemistry and π-systems were standardized, and charges were neutralized. The resulting set contained 512 kinases and 123,246 bioactivities. For training and validation of the classification models the threshold for 'active' compounds was set at $pChEMBL/pK_i/pIC_{50}/pEC_{50} \geq 6.5$ as we did previously.[37] Compounds that did not reach this threshold were termed 'inactive'. All compounds per target were divided into five subsets using clustering with the Cluster Molecules component (FCFP4 clustering) in Pipeline Pilot 2016. Compounds were clustered into five clusters per each kinase family (same L2 level in ChEMBL). First, compounds were separated based on activity (active when pChEMBL > 6.5) and then clustered into five clusters, after which active and inactive compounds were combined again. To ensure that every panel kinase was represented in each cluster, clustering into five clusters was done in two steps: the panel kinase was clustered first, followed by co-clustering of compounds from related kinases into the same cluster. The biggest cluster, or subset, was used as fixed training set, while the other four subsets were rotationally used as test or training set in 4-fold cross-validation. Datasets used in QSAR modeling only contained information of the respective kinase. In PCM modeling, additionally data of related kinases was included based on L4 classification in ChEMBL.

Dataset structure-based docking

For each kinase target, with the exception of PAK3 and ERK8 because of lack of a crystal structure, a benchmark set was created to validate structure-based docking for each target. The benchmarking set for structure-based docking included active compounds, inactive compounds, and decoy compounds, which were derived from ChEMBL (version 23) and the

DUD-e webserver[15]. Compounds with reported pChEMBL value for the challenge kinases, with confidence score 9, assay type B, and molecular weight < 550, were standardized using BIOVIA Pipeline Pilot 2016[36] by removing salts and keeping the largest fragment. An activity gap was realized to better distinguish active and inactive compounds: compounds with a pChEMBL value > 6.5 were assigned the label 'active', and compounds with a pChEMBL value > 4 and < 5.5 were labeled 'inactive'. Compounds with pChEMBL value ≤ 4 were excluded as inactive compounds to limit the number of inactives. Exceptions in the thresholds were made for activity labeling of compounds for targets RET and MKNK1: for RET compounds were labeled as inactive when pChEMBL value > 4.5 and < 6.5, and for MKNK1 compounds were defined as active when pChEMBL value > 6 to increase the fraction of active compounds. The thresholds used for RET were not intended to alter the number of (in)active compounds; RET was used as exploratory case and therefore for this kinase 'initial' thresholds were used. However, the 'general' threshold for all kinases was adapted since this yielded increased performance.

Scaffold trees were generated for the compounds based on Bemis-Murcko scaffolds[26] and the compound with the highest pChEMBL value per scaffold class, per activity class, per kinase, was kept. Compounds labeled as active were clustered into 100 clusters using the Cluster Molecules component (k-means) in BIOVIA Pipeline Pilot 2016. From the resulting clusters, only the cluster centers were kept. Kinases that had less than 100 active compounds available were excluded from this clustering step. The resulting active compound sets of maximum 100 compounds per kinase were used to generate decoys for each target kinase using the DUD-e webserver. For target PIK3CA decoys were generated based on only 80% of the active molecules as DUD-e decoy generation failed for the remaining fraction. The decoys were considered as inactive compounds when used in model validation. The collected actives, inactives, and decoys were prepared for docking using LigPrep from Schrödinger[23].

Screening dataset

Compounds for virtual screening were extracted from the ZINC15 compound library[16]. These compounds were selected based on "in stock" status and were filtered on a maximum of one Lipinski's rule of five violation. Additionally, the screening set was filtered on novelty by only including compounds with Tanimoto similarity (ECFP6) < 0.4 compared to known 'active' compounds (pChEMBL value ≥ 5) in ChEMBL (version 23) for the kinases of the respective panel. For panel 2, covering AURKA, PAK1, FGFR1, and LKB1, PAK3, TAK1, and PIK3CA, the screening set was additionally filtered on PSA < 75 to allow permeability of the blood-brain barrier. The compounds in the screening dataset were additionally prepared for structure-based docking by using LigPrep.

Statistical modeling - QSAR

Single target QSAR models were constructed for the main on-target kinases RET, AURKA, and PAK1. QSARs were built using BIOVIA Pipeline Pilot 2016 (version 18.0.1.1604). Models were trained on compound descriptors FCFP4 (3000 most frequent bits) and 86 physicochemical descriptors (**S1**). The following settings were applied in training categorical Random Forest QSAR models: 1000 trees, log2(m) number of descriptors, equalized class sizes and seed 12345. The classification threshold for active compounds was set at pChEMBL > 6.5.

Statistical modeling - PCM

PCM modeling was performed on all 512 kinases in the statistical modeling dataset. A multi-target PCM model was constructed using a random forest classifier in scikit-learn[38]. Compound descriptors were the same as used in QSAR modeling. Protein descriptors were calculated based on full sequence alignment derived from kinase.com (with a total of 1567 alignment positions including gaps).[39,40] The residues in the alignment were converted to three z-scales and an additional mean value for each of the three z-scales was added per sequence, resulting in a total of 4704 protein descriptors per kinase. Gaps were included in these descriptors, and were assigned a value of "0" for all three z-scales as was done previously.[41,42]

The random forest model was of high complexity due to the high dimensionality of the data, which includes the compounds' physiochemical properties, FCFP4 descriptors, and protein descriptors. The complexity of the random forest models was reduced by imposing constraints on parameters, such as the number of trees and maximum depth of the trees. The hyperparameters of the random forest model were optimized by utilizing random search, which evaluates the performance of the algorithm using different and randomly chosen configurations with cross-validation. Random search can be considered a simple form of Automated Machine Learning[19,43] and has shown to outperform other basic methods of hyperparameter optimization, such as grid search.[44] Approximately 500 random configurations were evaluated, resulting in an improvement of on average 7% AUROC over the default configuration (300 trees, log2(m) features at every node in the tree, no maximum tree depth, a minimum of 2 samples to consider a node for splitting, and bootstrap enabled) in 4-fold cross-validation. The settings that corresponded with the best model in random parameter optimization were applied in model training on the full data set, which was used in virtual screening. The final model consisted of 300 trees (fixed) with 43 features at every node in the tree, a maximum tree depth of 99, a minimum of 12 samples to consider a node for splitting, and bootstrapping disabled.

Structure-based docking

X-ray structures of all challenge proteins were extracted from the PDB[45] (except for ERK8 and PAK3 as no structure was available at the time). Crystal structures lacking co-crystalized orthosteric small-molecule ligands were discarded. The remaining structures were prepared for docking with the Protein Preparation tool from the Schrödinger 2017-4 suite after removing any other components than the protein, orthosteric ligand, and binding pocket ions. The 'add missing side chains' option was used, waters were removed, hydrogens were added, and disulphide bonds were created. The crystal structures were superposed per target and co-crystalized ligands were removed from the binding site. The grid for docking was determined for each target by the center of one of the co-crystalized ligands (box size xyz = 35 Å). Compounds were docked into the binding pocket using the Schrödinger 2017-4 suite[23] and the OPLS3 force field[46] with standard precision (SP) and standard settings. A maximum of ten poses per compound (per target) was generated.

Induced-fit docking

Induced-fit docking, as implemented in the Schrödinger 2017-4 suite,[23] was applied to kinases LKB1, S6K, and MKNK1. These kinases showed poor enrichment of actives when docked into the available crystal structures (sum of ROC and BEDROC < 1). For S6K five active

compounds, for MKNK1 ten active compounds, and for LKB1 six active compounds were docked using induced-fit docking (see **Table S5.4** for list of compounds). The crystal structures selected for induced-fit docking were 2HW6 for MKNK1, 2WTK for LKB1, and 3WF5, 3A62, and 4RLO for S6K. The resulting protein conformations were used as alternative protein structures in addition to the original crystal structures.

Structural protein-ligand interaction fingerprints (SPLIFs)

Structural protein-ligand interaction fingerprints (SPLIFs)[47] were calculated for a maximum of ten poses per compound that were retrieved from structure-based docking. The co-crystalized ligands of the protein structures were used as reference in SPLIF calculations to derive Tanimoto-like SPLIF scores. Five protein structures per kinase were selected for SPLIF generation. These structures were selected based on diversity of their co-crystalized ligands and best active-enrichment based on docking scores. The diversity of the co-crystalized ligands was assessed using affinity propagation clustering based on FCFP6 similarity[6]. One protein structure was selected from every corresponding cluster based on best docking score performance until a maximum of five protein structures was reached. Subsequently, Tanimoto-like SPLIF scores were calculated for the compounds in the benchmark sets for each of the selected protein structures. For each compound (maximum 10 poses) the best SPLIF score was used to calculate actives enrichment based on SPLIF scores.

For RET and PIK3CA exceptions were made in the selection of benchmark proteins for SPLIF because only four and three clusters were generated, respectively. Therefore, similar co-crystalized ligands and their corresponding proteins were also selected to get a total of five protein structures per target. Furthermore, for RET structure 2IVV instead of the co-crystalized ligand, CHEMBL3775169 was used as a reference.

Binding pose metadynamics

A metadynamics-composition score[10] was calculated with the binding pose metadynamics tool in the Schrödinger 2017-4 suite,[23] for compounds and poses derived from structure-based docking. The top 100 compounds for RET (PDB: 2IVU) and AURKA (PDB: 2BMC) were selected, based on best docking scores. The protein structures were prepared for metadynamics simulations by capping the termini and the run time of each simulation was set at 10 ns. The resulting metadynamics-composition scores were added to the existing docking scores to re-rank the top 100 compounds in the benchmark set, and to re-rank the top 100 compounds in the screening set.

Ensemble scoring in structure-based modeling

For every target, protein models for virtual screening were selected based on ensemble scoring of docking and SPLIF scores. The combination, or ensemble, of protein models that resulted in the best actives enrichment was used in virtual screening. The validated ensembles were based on docking scores of the top five enriching models and all five SPLIF models per target. Different ensembles were tested for each target including: docking scores only, SPLIF scores only, and docking and SPLIF scores combined. The performances of the ensembles were evaluated using Z2-scoring by averaging over the top two Z-scores.[47] Prior to ensembling, the Z-scores per compound were normalized towards the actives from the respective benchmark set. Docking

scores were normalized to Z-scores by subtracting the mean docking score (mean over docking scores from all actives in the benchmark set) from the docking score of the test compound, and subsequently dividing by the standard deviation of the (benchmark) actives' docking scores. The same approach was used in normalization of SPLIF scores to Z-scores. However, the SPLIF scores were first multiplied by -1 to change the vector into the same direction as the docking scores (better binder, more negative score).

Compound ranking per target

The compounds in virtual screening were ranked for ensembles of targets based on the scores derived from PCM modeling and structure-based ensemble score. The PCM predictions were given a confidence weight between 0 and 1 based on the prediction performance per target (with 0 being no confidence, and 1 being highest confidence). The confidence score for the PCM model could be calculated using the following equation:

$$confidence\ weight_{PCM} = (ROC * 2 - 1) * \frac{\sqrt{nr\ of\ samples}}{\sqrt{\max nr\ of\ samples}}$$

The PCM confidence score was based on the number of compounds (of the corresponding target) in training and the ROC score derived from 4-fold cross-validation: ROC * 2 - 1 (to normalize that 0.5 corresponds with a weight of 0 and 1 corresponds with a weight of 1), multiplied by the weight based on the number of samples for the target in the training data (calculated as sqrt(number of samples)/sqrt(maximum number of samples).

The weights of the structure-based predictions were calculated by taking the total sum of ROC and BEDROC per target[48], consequently giving the structure-based predictions more weight than the statistical PCM model predictions. However, structure-based models were penalized (by multiplying with 0.5) when induced-fit models were created for the kinase. Moreover, kinases containing less than 100 actives in the benchmark set were penalized by multiplying the structure-based weight with the fraction of number of actives (with a fraction of 1 being 100 actives, and a fraction of 0.1 being 10 actives).

Finally, the obtained weights for PCM and structure-based models were applied in calculating the compound rank per kinase.

$$\begin{aligned} compound\ rank \\ &= confidence\ weight_{PCM} * active\ class\ probability \\ &+ confidence\ weight_{structure-based} * -Z2\ score \end{aligned}$$

Compound predictions were multiplied by the subtotal weights and summed up (statistical model weight * statistical PCM prediction + structure-based model weight * structure-based Z2 score), resulting in a final compound rank per target.

Protein kinase assay

A selection of 46 compounds was experimentally validated for inhibitory activity against RET. The compounds were ordered via Mcule Inc.[49], (Budapest) and purchased from ChemBridge, ChemDiv, ChemScene, Enamine, Life Chemicals, and Vitas M Chemical Limited. The assays were performed by ProQinase GmbH[50], Germany. First, the compounds were tested (n=1) at final assay concentrations 10 μM and 0.1 μM. The five resulting hits (< 60% remaining RET activity at concentration 10 μM) were re-evaluated by determination of IC_{50} values (n=3). RET WT activity was measured using a radiometric protein kinase assay (^{33}PanQinase® Activity Assay). All kinase assays were performed in 96-well FlashPlatesTM from PerkinElmer (Boston, MA, USA) in a 50 ml reaction volume. The reaction cocktail was pipetted in 4 steps in the following order: 1) 20 ml of assay buffer. 2) 5 ml of ATP solution (in H_2O). 3) 5 ml of test compound (in 10 % DMSO), and 4) 20 μl enzyme/substrate mix. The assay for all protein kinases contained 70 mM HEPES-NaOH pH 7.5, 3 mM $MgCl_2$, 3 mM $MnCl_2$, 3 mM Na-orthovanadate, 1.2 mM DTT, 50 μg/ml PEG20000, 1 μM ATP, [γ -^{33}P]-ATP (approx. 1.91 x 1005 cpm per well), 40 ng/50 μl protein kinase, and 0.125 μg/50μl poly(Glu, Tyr)4:1 substrate. The reaction cocktails were incubated at 30ºC for 60 minutes. The reaction was stopped with 50 ml of 2 % (v/v) H_3PO_4, plates were aspirated and washed two times with 200 ml 0.9 % (w/v) NaCl. Incorporation of ^{33}Pi was determined with a microplate scintillation counter (Microbeta, Wallac).

References

(1) de Lera, A. R.; Ganesan, A. Epigenetic Polypharmacology: From Combination Therapy to Multitargeted Drugs. Clin. Epigenetics 2016, 8, 105. https://doi.org/10.1186/s13148-016-0271-9.
(2) van Leeuwen, R. W. F.; Jansman, F. G. A.; van den Bemt, P. M. L. A.; de Man, F.; Piran, F.; Vincenten, I.; Jager, A.; Rijneveld, A. W.; Brugma, J. D.; Mathijssen, R. H. J.; van Gelder, T. Drug–Drug Interactions in Patients Treated for Cancer: A Prospective Study on Clinical Interventions†. Ann. Oncol. 2015, 26, 992–997. https://doi.org/10.1093/annonc/mdv029.
(3) Schlessinger, A.; Abagyan, R.; Carlson, H. A.; Dang, K. K.; Guinney, J.; Cagan, R. L. Multi-Targeting Drug Community Challenge. Cell Chem. Biol. 2017, 24, 1434–1435. https://doi.org/https://doi.org/10.1016/j.chembiol.2017.12.006.
(4) Dar, A. C.; Das, T. K.; Shokat, K. M.; Cagan, R. L. Chemical Genetic Discovery of Targets and Anti-Targets for Cancer Polypharmacology. Nature 2012, 486, 80.
(5) Sonoshita, M.; Scopton, A. P.; Ung, P. M. U.; Murray, M. A.; Silber, L.; Maldonado, A. Y.; Real, A.; Schlessinger, A.; Cagan, R. L.; Dar, A. C. A Whole-Animal Platform to Advance a Clinical Kinase Inhibitor into New Disease Space. Nat. Chem. Biol. 2018, 14, 291.
(6) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742–754. https://doi.org/10.1021/ci100050t.
(7) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. Med. Chem. Commun. 2011, 2, 16–30. https://doi.org/10.1039/C0MD00165A.
(8) Burggraaff, L.; Oranje, P.; Gouka, R.; van der Pijl, P.; Geldof, M.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 Using Proteochemometric Modeling. J. Cheminform. 2019, 11, 15. https://doi.org/10.1186/s13321-019-0337-8.
(9) Christmann-Franck, S.; van Westen, G. J. P.; Papadatos, G.; Beltran Escudie, F.; Roberts, A.; Overington, J. P.; Domine, D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? J. Chem. Inf. Model. 2016, 56, 1654–1675. https://doi.org/10.1021/acs.jcim.6b00122.
(10) Clark, A. J.; Tiwary, P.; Borrelli, K.; Feng, S.; Miller, E. B.; Abel, R.; Friesner, R. A.; Berne, B. J. Prediction of Protein–Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. J. Chem. Theory Comput. 2016, 12, 2990–2998. https://doi.org/10.1021/acs.jctc.6b00201.
(11) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. Nucleic Acids Res. 2017, 45, D945–D954. https://doi.org/10.1093/nar/gkw1074.
(12) Eidogen Sertanty http://www.eidogen-sertanty.com/ (accessed Jan 22, 2018).
(13) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. J. Cheminform. 2017, 9, 17. https://doi.org/10.1186/s13321-017-0203-5.
(14) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. J. Comput. Aided. Mol. Des. 2015, 29, 885–896. https://doi.org/10.1007/s10822-015-9860-5.
(15) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J. Med. Chem. 2012, 55, 6582–6594. https://doi.org/10.1021/jm300687e.
(16) Sterling, T.; Irwin, J. J. ZINC 15--Ligand Discovery for Everyone. J. Chem. Inf. Model. 2015, 55, 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.
(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 2001, 46, 3–26.
(18) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. Nucleic Acids Res. 2014, 42, D1083-90. https://doi.org/10.1093/nar/gkt1031.
(19) Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res. 2012, 13, 281–305.
(20) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. J. Chem. Inf. Model. 2014, 54, 2555–2561. https://doi.org/10.1021/ci500319f.
(21) Frey, B. J.; Dueck, D. Clustering by Passing Messages Between Data Points. Science (80-.). 2007, 315, 972 LP – 976. https://doi.org/10.1126/science.1136800.
(22) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. J. Chem. Inf. Model. 2013, 53, 1531–1542. https://doi.org/10.1021/ci300463g.
(23) Schrödinger. Schrödinger Maestro Release 2017-4. Schrödinger: LLC, New York, NY 2017.

(24) Tiwary, P.; Parrinello, M. A Time-Independent Free Energy Estimator for Metadynamics. J. Phys. Chem. B 2015, 119, 736–742. https://doi.org/10.1021/jp504920s.
(25) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. Nucleic Acids Res. 2016, 44, D1220-8. https://doi.org/10.1093/nar/gkv1253.
(26) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. 1996, 39, 2887–2893. https://doi.org/10.1021/jm9602928.
(27) Atzori, A.; Bruce, N. J.; Burusco, K. K.; Wroblowski, B.; Bonnet, P.; Bryce, R. A. Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics. J. Chem. Inf. Model. 2014, 54, 2764–2775. https://doi.org/10.1021/ci5003334.
(28) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. Drug Discov. Today 2007, 12, 225–233. https://doi.org/https://doi.org/10.1016/j.drudis.2007.01.011.
(29) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. J. Chem. Inf. Model. 2005, 45, 839–849. https://doi.org/10.1021/ci0500381.
(30) Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-House HTS Data. ACS Chem. Biol. 2016, 11, 1255–1264. https://doi.org/10.1021/acschembio.6b00029.
(31) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. Nature 2019, 566, 224–229. https://doi.org/10.1038/s41586-019-0917-9.
(32) Liu, Y.; Gray, N. S. Rational Design of Inhibitors That Bind to Inactive Kinase Conformations. Nat. Chem. Biol. 2006, 2, 358–364. https://doi.org/10.1038/nchembio799.
(33) Vijayan, R. S. K.; He, P.; Modi, V.; Duong-Ly, K. C.; Ma, H.; Peterson, J. R.; Dunbrack Jr, R. L.; Levy, R. M. Conformational Analysis of the DFG-out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. J. Med. Chem. 2015, 58, 466–479. https://doi.org/10.1021/jm501603h.
(34) Miljkovic, F.; Rodríguez-Pérez, R.; Bajorath, J. Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes. J. Med. Chem. 2019. https://doi.org/10.1021/acs.jmedchem.9b00867.
(35) van Westen, G. J. P.; Gaulton, A.; Overington, J. P. Chemical, Target, and Bioactive Properties of Allosteric Modulation. PLoS Comput. Biol. 2014, 10, e1003559. https://doi.org/10.1371/journal.pcbi.1003559.
(36) Dassault Systèmes BIOVIA. BIOVIA Pipeline Pilot. Dassault Systèmes: San Diego 2016.
(37) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using A ChEMBL Bioactivity Benchmark Set. bioRxiv 2017.
(38) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; VanderPlas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. CoRR 2012, abs/1201.0490.
(39) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. Science (80-.). 2002, 298, 1912 LP – 1934. https://doi.org/10.1126/science.1075762.
(40) kinase.com http://kinase.com (accessed May 8, 2017).
(41) van Westen, G. J.; Swier, R. F.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. J. Cheminform. 2013, 5, 41. https://doi.org/10.1186/1758-2946-5-41.
(42) van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. J. Cheminform. 2013, 5, 42. https://doi.org/10.1186/1758-2946-5-42.
(43) Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning. In Advances in Neural Information Processing Systems 28; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2962–2970.
(44) Kotthoff, L.; Thornton, C.; Hoos, H. H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA. J. Mach. Learn. Res. 2017, 18, 1–5.
(45) Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Res. 2000, 235–242. https://doi.org/10.1093/nar/28.1.235.
(46) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J. Chem. Theory Comput. 2016, 12, 281–296. https://doi.org/10.1021/acs.jctc.5b00864.
(47) Lenselink, E. B.; Jespers, W.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. Interacting with GPCRs:

Using Interaction Fingerprints for Virtual Screening. J. Chem. Inf. Model. 2016, 56, 2053–2060. https://doi.org/10.1021/acs.jcim.6b00314.
(48) and, J.-F. T.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. 2007. https://doi.org/10.1021/CI600426E.
(49) mcule https://mcule.com (accessed Jun 13, 2019).
(50) ProQinase GmbH. ProQinase https://www.proqinase.com (accessed Aug 22, 2019).

Supporting Information

Table S5.1. List of 86 physicochemical descriptors.

Physicochemical descriptors			
CMP_ATOMCOUNT_C	CMP_NUM_CHAINASSEM-BLIES	CMP_ATOMCOUNT_I	CMP_ATOMS_O_FRAC
CMP_ATOMCOUNT_C_sp3	CMP_ATOMCOUNT_META-LATOMS	CMP_Molecular_solubility	CMP_ATOMS_HETEROATOM_ FRAC
CMP_ATOMCOUNT_C_sp2	CMP_BONDS_PIBONDS	CMP_LogP	CMP_ATOMS_S_FRAC
CMP_ATOMCOUNT_C_sp	CMP_BONDS_STEREOATOMS	CMP_MOLECULAR_SUR-FACEAREA	CMP_ATOMS_P_FRAC
CMP_ATOMS_C_sp3_FRAC	CMP_BONDS_STEREOBONDS	CMP_MOLECULAR_SASA	CMP_ATOMS_HALOGEN_ FRAC
CMP_ATOMS_C_sp2_FRAC	CMP_BONDS_SINGLEBONDS	CMP_MOLECULAR_POLAR-SASA	CMP_ATOMS_OTHER_FRAC
CMP_ATOMS_C_sp_FRAC	CMP_BONDS_DOUBLEBONDS	CMP_NUM_ALIPHATICRINGS	CMP_ATOMS_STEROATOM_ FRAC
CMP_FORMALCHARGE	CMP_BONDS_TRIPLEBONDS	CMP_BONDS_HEAVYATOMS	CMP_ATOMS_POSITIVE_FRAC
CMP_ATOMCOUNT_HEAVY	CMP_NUM_ALIPHATICSIN-GLEBONDS	CMP_BONDS_HYDROGENS	CMP_ATOMS_NEGATIVE_ FRAC
CMP_BONDS_TOTAL	CMP_NUM_ALIPHATICDOU-BLEBONDS	CMP_BONDS_SINGLE_FRAC	CMP_ATOMS_H_ACCEP-TOR_FRAC
CMP_ATOMCOUNT_HYDRO-GENS	CMP_NUM_HYDROGENBONS	CMP_BONDS_DOUBLE_FRAC	CMP_ATOMS_H_DONOR_ FRAC
CMP_BONDS_EXPLICIT	CMP_NUM_TERMINAL-ROTOMERS	CMP_BONDS_TRIPLE_FRAC	CMP_RIGIDITY_INDEX
CMP_ATOMCOUNT_POSITIVE	CMP_ATOMCOUNT_H_AC-CEPTORS	CMP_BONDS_AROMATIC_ FRAC	CMP_MOLECULAR_WEIGHT
CMP_ATOMCOUNT_NEGA-TIVE	CMP_ATOMCOUNT_H_DO-NORS	CMP_BONDS_BRIDGE_FRAC	CMP_CLASS_LIPINSKI_PASS
CMP_BONDS_RING	CMP_ATOMCOUNT_H	CMP_BONDS_STEREO_FRAC	CMP_NUM_Ro5_Violations
CMP_BONDS_ROTATABLE	CMP_ATOMCOUNT_N	CMP_BONDS_RING_FRAC	CMP_RO3_PASS
CMP_BONDS_AROMATIC	CMP_ATOMCOUNT_O	CMP_BONDS_ALIPHATIC_ FRAC	CMP_MOLECULAR_POLAR_ SURFACEAREA
CMP_BONDS_BRIDGE	CMP_ATOMCOUNT_F	CMP_BONDS_ROTATABLE_ FRAC	CMP_MOLECULAR_POLAR_ SURFACEAREA_FRAC
CMP_NUM_RINGS	CMP_ATOMCOUNT_P	CMP_ATOMCOUNT_HALO-GENS	CMP_FULL_MWT
CMP_NUM_AROMATICRINGS	CMP_ATOMCOUNT_S	CMP_ATOMS_C_FRAC	CMP_INORGANIC_FLAG
CMP_NUM_RINGASSEMBLIES	CMP_ATOMCOUNT_Cl	CMP_ATOMS_H_FRAC	
CMP_NUM_CHAINS	CMP_ATOMCOUNT_Br	CMP_ATOMS_N_FRAC	

Table S5.2. Selected structure-based models per target.

Protein	PDB	Model	BEDROC (α=160.9)	ROC	Sum
Challenge 1					
RET	2IVU	Docking+SPLIF	0.56	0.85	1.41
	5FM2	Docking			
	2IVV*	Docking+SPLIF			
	5FM3	Docking			
	2X2K	Docking			
BRAF	5HID	Docking+SPLIF	0.75	0.9	1.66
	1UWH	Docking			
	4DBN	Docking			
	4G9R	SPLIF			
SRC	4MXX	Docking+SPLIF	0.64	0.92	1.56
	2BDJ	Docking			
	1YOL	Docking			
	2H8H	SPLIF			
S6K	3A62 IFD #74	SPLIF	0.24	0.72	0.96
MKNK1	2HW6 IFD #25	SPLIF	0.71	0.86	1.57
PDK1	3QD0	Docking	0.66	0.87	1.53
TTK	4C4I	Docking	0.81	0.98	1.79
	5N84	Docking			
	5N87	SPLIF			
	5NA0	SPLIF			
Challenge 2					
AURKA	2BMC	Docking	0.52	0.82	1.34
	4B0G	SPLIF			
PAK1	4EQC	Docking+SPLIF	0.97	0.98	1.95
	6B16	SPLIF			
FGFR1	3GQL	Docking	0.68	0.94	1.62
	4RWL	Docking			
	3TT0	Docking			
	4NKA	Docking			
LKB1	2WTK IFD #83	Docking+SPLIF	0.67	0.81	1.48
	2WTK IFD #34	Docking+SPLIF			
TAK1	4L53	Docking+SPLIF	0.64	0.76	1.40
	4L3P	Docking			
	5JGB	SPLIF			
PIK3CA	4JPS	Docking	0.99	0.62	1.61
	5DXH	SPLIF			

Table S5.3. Bioactivities of 46 compounds validated for RET activity.

ZINC ID	Smiles	Selection method	Remaining RET activity with compound concentration 10^{-7}	Remaining RET activity with compound concentration 10^{-5}	IC_{50} (M) n1	IC_{50} (M) n2	IC_{50} (M) n3
ZINC33008650	COc1cc(NC(=O)Nc2ccc(C)c(F)c2)cc3sc(NC(=O)c4ccccc4)nc13	Consensus	107	16	5.16E-06	9.12E-06	7.87E-06
ZINC12324934	CCN1C=C(C(=O)Nc2cccc(O)c2)C(=O)c3cc(F)c(cc13)N4CCN(CC4)C(=O)C	Structure-based	91	25	1.65E-05	4.06E-05	1.81E-05
ZINC9518200	NC1=C(NC(=O)c2ccc(F)c(F)c2)C(=O)NC(=N1)SCC(=O)Nc3ccc4OCOc4c3	Structure-based	109	28	5.85E-05	7.30E-05	1.71E-04
ZINC72312837	COc1ccc(NC(=O)Nc2ccccc2c3nc(n[nH]3)c4cocc4)cc1	Consensus	103	54	1.99E-05	1.74E-05	3.67E-05

ZINC ID	Smiles	Selection method	Remaining RET activity with compound concentration 10^{-7}	Remaining RET activity with compound concentration 10^{-5}
ZINC65184824	COc1cccc(Nc2nc(C)c(s2)C(=O)Nc3ccc4OCOc4c3)c1	Structure-based	129	55
ZINC101094323	Cc1cc(NC(=O)NC2=C(NC3CCCC3)c4ccccc4OC2=O)ccc1Br	Structure-based	97	68
ZINC72119413	Cc1ccc(NC(=O)Nc2ccc(Nc3cc(nc(C)n3)n4ccnc4)cc2)c(C)c1	Consensus	95	68
ZINC33008371	Cc1cc(NC(=O)Nc2ccc(F)cc2)cc3sc(NC(=O)c4ccccc4)nc13	Consensus	146	79
ZINC49407610	COC1=CC(=O)C(=NN1c2ccc(OC)c(OC)c2)C(=O)Nc3ccc4OCOc4c3	Structure-based	120	79
ZINC65250057	Cc1cccc(c1)n2ncc3C(=C(C(=O)Nc4ccc(C)cc4F)C(=O)Nc23)O	Structure-based	134	79
ZINC33008195	COc1ccccc1NC(=O)Nc2ccc3nc(NC(=O)c4ccccc4)sc3c2	Consensus	113	84
ZINC33008629	CCOc1ccc(NC(=O)Nc2cc(OC)c3nc(NC(=O)c4ccccc4)sc3c2)cc1	Consensus	115	85
ZINC101332769	Fc1ccc(NC(=O)CN2C(=O)N(CC3CCCO3)C(=O)c4ccc(cc24)C(=O)NCCc5ccccc5)cc1Cl	Structure-based	116	85

ZINC ID	Smiles	Selection method	Remaining RET activity with compound concentration 10^{-7}	Remaining RET activity with compound concentration 10^{-5}
ZINC12584529	Cc1cccc(NC(=O)CNC(=O)c2ccccc2NS(=O)(=O)c3ccc4NC(=O)Nc4c3)c1C	Structure-based	114	86
ZINC33008381	COc1cc(NC(=O)Nc2cc(C)c3nc(NC(=O)c4ccccc4)sc3c2)cc(OC)c1	Consensus	123	88
ZINC6864637	O=C1Nc2ccc(NC=C3C(=O)NC(=O)N(C3=O)c4ccccc4)cc2N1	Structure-based	107	91
ZINC67956408	O=C(NCc1cccnc1n2cnc3ccccc23)c4ccc5NC(=O)Nc5c4	Structure-based	102	91
ZINC83815577	CC(=O)Nc1cccc2CN(CCc12)C(=O)CCNC(=O)c3ccsc3	Structure-based	104	91
ZINC21806194	CCc1cccc(NC(=O)c2cc([nH]n2)c3ccc(OC)c(c3)S(=O)(=O)N)c1	Structure-based	120	93
ZINC9784888	CC(=O)c1cccc(NC(=O)Cc2csc(NC(=O)Nc3cccc(Cl)c3)n2)c1	Structure-based	116	93
ZINC33008624	COc1cc(NC(=O)Nc2ccccc2Br)cc3sc(NC(=O)c4ccccc4)nc13	Statistical modelling	106	93
ZINC33008616	COc1ccc(OC)c(NC(=O)Nc2cc(OC)c3nc(NC(=O)c4ccccc4)sc3c2)c1	Consensus	114	94
ZINC49002135	Cc1ccc2C(=O)C(=CN(CC(=O)NC3CCCC3)c2n1)C(=O)Nc4ccc(C)c(C)c4	Structure-based	114	96
ZINC35868136	COc1cccc(NC(=O)Nc2c(N)n(nc2c3ccc(Cl)cc3)c4ccc(F)cc4)c1	Statistical modelling	110	96
ZINC33008380	COc1ccc(NC(=O)Nc2cc(C)c3nc(NC(=O)c4ccccc4)sc3c2)c(OC)c1	Consensus	133	99
ZINC1430237	NC1=Nc2nc3ccccc3n2C(N1)c4c[nH]nc4c5cccc(O)c5	Structure-based	117	99
ZINC46009367	CN1C(=O)N(C)c2cc(Oc3ccccc3)c(NC(=O)Nc4cccc(c4)C(=O)C)cc12	Consensus	104	100
ZINC23135952	C1(N2CCN(CC2)C(=O)C2=CC=CO2)N=C2NN=CC2=C(NC2=CC=C(C=C2)C)N=1	Structure-based	130	100
ZINC12155670	COCC(=O)Nc1cc(cc2ncn(CCc3ccccc3)c12)C(=O)NCc4ccc(C)c(F)c4	Structure-based	133	100
ZINC103927245	FC(F)(F)c1cc(NC(=O)c2cocc2)ccc1NC(=O)Nc3ccc(NC(=O)c4cocc4)cc3C(F)(F)F	Statistical modelling	113	101
ZINC33008656	COc1cc(NC(=O)Nc2cc(OC)c3nc(NC(=O)c4ccccc4)sc3c2)cc(OC)c1	Consensus	112	101
ZINC33008626	CCOc1ccccc1NC(=O)Nc2cc(OC)c3nc(NC(=O)c4ccccc4)sc3c2	Consensus	118	101
ZINC2955974	CC(=O)OCc1cnc(C)c2OC(=Nc3ccc(F)cc3)C(=Cc12)C(=O)Nc4ccc5OCOc5c4	Structure-based	121	101
ZINC33008219	COc1ccc(Cl)cc1NC(=O)Nc2ccc3nc(NC(=O)c4ccccc4)sc3c2	Consensus	141	103

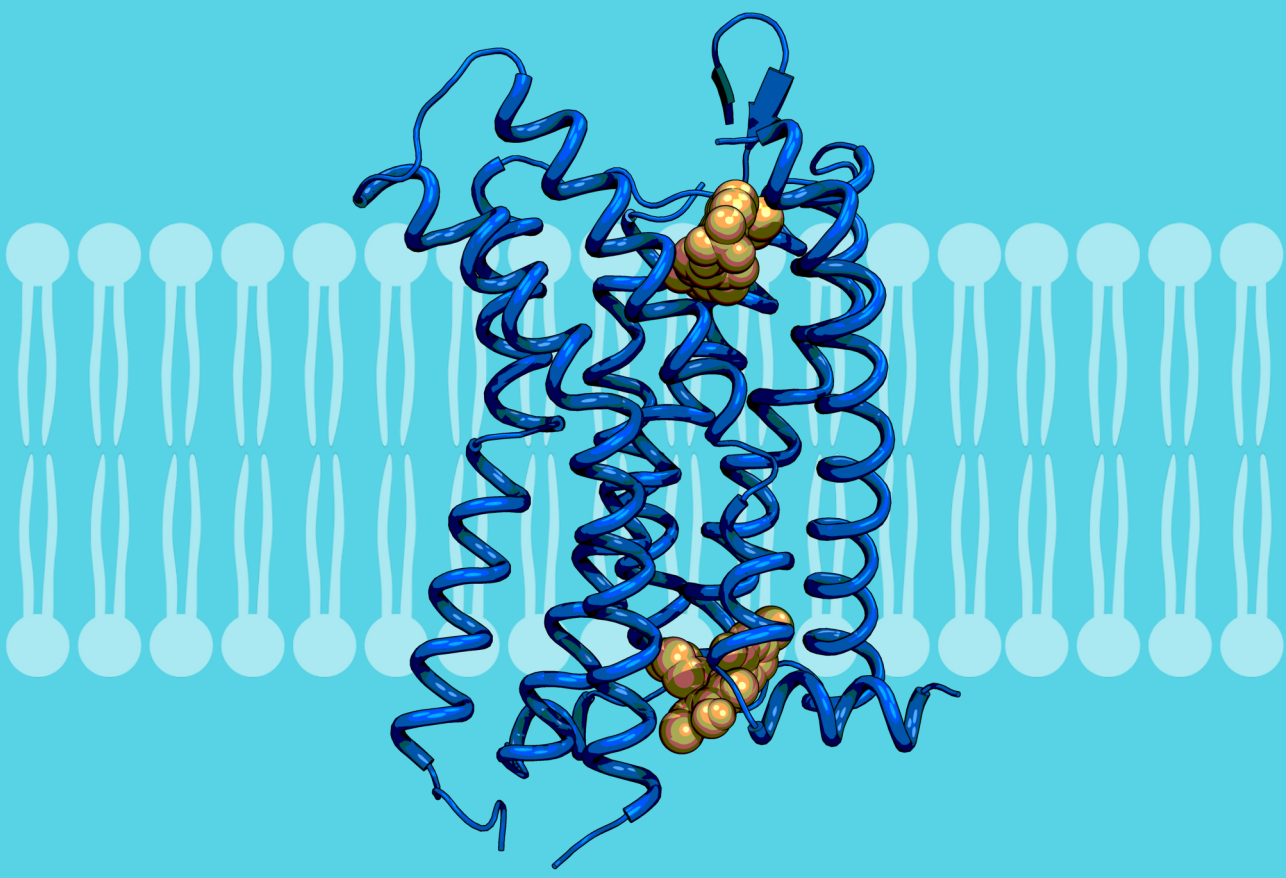
ZINC ID	Smiles	Selection method	Remaining RET activity with compound concentration 10^{-7}	Remaining RET activity with compound concentration 10^{-5}
ZINC33008399	COC(=O)c1cccc(NC(=O)Nc2cc(C)c3nc(NC(=O)c4ccccc4)sc3c2)c1	Consensus	109	103
ZINC103832769	FC(F)(F)c1cc(NC(=O)c2occc2)ccc1NC(=O)Nc3ccc(NC(=O)c4occc4)cc3C(F)(F)F	Statistical modelling	123	103
ZINC32996009	Cc1ccc(cc1)C2=NN3C(=O)C=C(CNC(=O)Nc4ccc(C)c(Cl)c4)N=C3S2	Structure-based	116	105
ZINC33008637	COc1cc(NC(=O)Nc2ccc(Br)cc2)cc3sc(NC(=O)c4ccccc4)nc13	Statistical modelling	120	106
ZINC12243694	COCC(=O)Nc1cc(cc2ncn(CCc3ccccc3)c12)C(=O)NCCc4cnn(C)c4	Structure-based	118	109
ZINC1211467	O=C(Cn1cc(C=C2C(=O)NC(=O)NC2=O)c3ccccc13)Nc4ccccc4	Structure-based	131	110
ZINC10279818	Clc1ccc(c(NC(=O)COc2ccc(Cl)cc2C(=O)Nc3ccccc3)c1)n4cncn4	Structure-based	107	111
ZINC72429138	COc1ccc(C2=Cn3cc(nc3C(=O)N2C)C(=O)N4CC(O)C4)c(F)c1	Structure-based	117	111
ZINC11966082	COCC(=O)Nc1cc(C(=O)N(C)CCc2c(C)n[nH]c2C)c3c(c1)ncn3Cc4ccc(F)cc4	Structure-based	118	117
ZINC33008631	COc1cc(NC(=O)Nc2cccc(Br)c2)cc3sc(NC(=O)c4ccccc4)nc13	Statistical modelling	153	127
ZINC46009379	CN1C(=O)N(C)c2cc(Oc3ccccc3)c(NC(=O)Nc4ccc(C)c(C)c4)cc12	Consensus	112	128
ZINC62022653	C1(N2CCN(CC2)C(OCC)=O)N=C(NC2C=CC(=CC=2)C)C2=C(NN=N2)N=1	Structure-based	128	154

Table S5.4. List of compounds used to generate structures for S6K, MKNK1, and LKB1, with induced-fit docking.

Target kinase	Compound ID	Compound SMILES
S6K	CHEMBL6291	Cn1cc(C2=C(C(=O)NC2=O)c3cn(CCCSC(=N)N)c4ccccc34)c5ccccc15
S6K	CHEMBL3685220	NC(=O)c1cccc2c(NCc3cccc(c3)N4C(=O)Nc5ccccc5C4=O)ncnc12
S6K	CHEMBL3680430	COCCC(Nc1ncnc2c(cccc12)C(=O)N)c3cccc(NC(=O)c4cc(n[nH]4)C5CC5)c3
S6K	CHEMBL3685220	NC(=O)c1cccc2c(NCc3cccc(c3)N4C(=O)Nc5ccccc5C4=O)ncnc12
S6K	CHEMBL3675398	CNCC(Nc1ncnc2c(cccc12)C(=O)N)c3cccc(NC(=O)c4ccc(F)c(F)c4)c3
MKNK1	CHEMBL3355002	CC(C)Oc1cc2[nH]ncc2cc1Nc3ncnc4[nH]c5CCC(Cc5c34)C(=O)NCCCS(=O)(=O)C
MKNK1	CHEMBL3355000	COc1cc2[nH]ncc2cc1Nc3ncnc4[nH]c5CCC(Cc5c34)C(=O)NCc6ccccc6
MKNK1	CHEMBL3354998	CC(C)Oc1cc2[nH]ncc2cc1Nc3ncnc4[nH]c5CCCCc5c34
MKNK1	CHEMBL223360	Cc1ccc(F)c(NC(=O)Nc2ccc(cc2)c3cccc4[nH]nc(N)c34)c1
MKNK1	CHEMBL3659492	Cc1c(sc2ncnc(Nc3ccc(F)cc3OC4CCN(CC(=O)N5CCOCC5)CC4)c12)C(=O)N
MKNK1	CHEMBL3417208	CN1C=C(Nc2ncnc3sc(C(=O)N)c(C)c23)C=CC1=O
MKNK1	CHEMBL1721885	Cc1[nH]c(\C=C\2/C(=O)Nc3ccc(F)cc23)c(C)c1C(=O)NC[C@H](O)CN4CCOCC4
MKNK1	CHEMBL226403	O=C1NCCc2[nH]c(cc12)c3ccnc(c3)c4cnc5ccccc5c4
MKNK1	CHEMBL3632739	COC(=O)c1sc2ncnc(Nc3ccc(F)cc3OC(C)C)c2c1C
MKNK1	CHEMBL3805029	FC(F)(F)c1cc(NC(=O)COc2ccc(cc2)c3cccnc3)ccc1CN4CCNCC4
LKB1	CHEMBL502835	COC(=O)c1ccc2\C(=C(\Nc3ccc(cc3)N(C)C(=O)CN4CCN(C)CC4)/c5ccccc5)\C(=O)Nc2c1
LKB1	CHEMBL603469	C[C@]12O[C@H](C[C@]1(O)CO)n3c4ccccc4c5c6C(=O)NCc6c7c8ccccc8n2c7c35
LKB1	CHEMBL1908397	O=C(N1CCNCC1)c2ccc(\C=C\c3n[nH]c4ccccc34)cc2
LKB1	CHEMBL535	CCN(CC)CCNC(=O)c1c(C)[nH]c(\C=C\2/C(=O)Nc3ccc(F)cc23)c1C
LKB1	CHEMBL1287853	Cc1cnc(Nc2ccc(OCCN3CCCC3)cc2)nc1Nc4cccc(c4)S(=O)(=O)NC(C)(C)C
LKB1	CHEMBL388978	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n3c4ccccc4c5c6CNC(=O)c6c7c8ccccc8n2c7c35

Chapter six

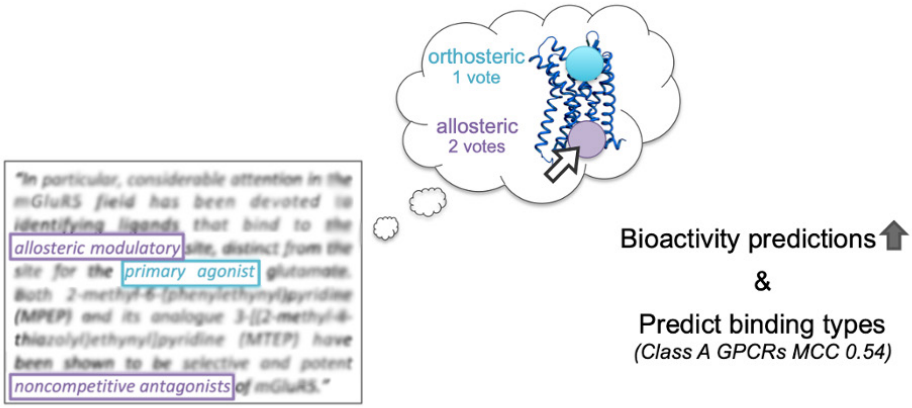
Annotation of Allosteric Compounds to Enhance Bioactivity Modeling for Class A GPCRs



Lindsey Burggraaff, Amber van Veen, Chi Chung Lam, Herman W.T. van Vlijmen, Adriaan P. IJzerman, and Gerard J.P. van Westen. *submitted*

Abstract

Proteins often have both orthosteric and allosteric binding sites. Endogenous ligands, such as hormones and neurotransmitters, bind to the orthosteric site, while synthetic ligands may bind to orthosteric or allosteric sites, which has become a focal point in drug discovery. Usually such allosteric modulators bind to a protein non-competitively with its endogenous ligand or substrate. The growing interest in allosteric modulators has resulted in a substantial increase of these entities and their features such as binding data in chemical libraries and databases. Although this data surge fuels research focused on allosteric modulators, binding data is unfortunately not always clearly indicated as being allosteric or orthosteric. Therefore, allosteric binding data is difficult to retrieve from databases that contain a mixture of allosteric and orthosteric compounds. This decreases model performance when statistical methods, such as machine learning models, are applied. In previous work we generated an allosteric data subset of ChEMBL release 14. In the current study an improved text mining approach is used to retrieve the allosteric and orthosteric binding types from the literature in ChEMBL release 22. Moreover, convolutional deep neural networks were constructed to predict the binding types of compounds for class A G protein-coupled receptors (GPCRs). Temporal split validation showed the model predictiveness with MCC = 0.54, sensitivity allosteric = 0.54, and sensitivity orthosteric = 0.94. Finally, this study shows that inclusion of accurate binding types increases binding predictions by including them as descriptor (MCC = 0.27 improved to MCC = 0.34; validated for class A GPCRs, trained on all GPCRs). Although the focus of this study is mainly on class A GPCRs, binding types for all protein classes in ChEMBL were obtained and explored.



Introduction

Drugs bind to proteins to exert a biological effect, which can be activation, deactivation, or blockage to prevent an endogenous ligand from binding. The location on the protein where the endogenous, or natural, ligand binds is called the orthosteric binding site. All other binding sites are described as allosteric binding sites.[1] Allosteric binding of ligands (allosteric modulation) has been reported and studied at many protein classes, including membrane-bound G protein-coupled receptors (GPCRs).[1–3] Moreover, structural information is available in the Protein DataBank,[4] which shows the binding of allosteric ligands to GPCRs at different sites (**Figure 6.1**).[2,5] The transmembrane (TM) domains of the orthosteric binding pocket in GPCRs are highly conserved for most GPCR classes. An exception is formed by class C GPCRs, for which the orthosteric site lies in the extracellular Venus flytrap domain. The most common allosteric binding site for class C GPCRs is located close to or at the orthosteric site in the TM domain of other (class A) GPCRs.[6] Other GPCR allosteric binding sites have been identified close to or in the intracellular domain and between helices in the TM domain.[2,3] The physicochemical properties of these multiple binding sites can vary greatly, enabling dissimilar ligands to bind the same protein via different sites.

Although allosteric modulators are widely studied and reported, allosterism of compounds is not yet annotated in the public chemical database ChEMBL.[7] Previous attempts have been made to retrieve allosteric properties of compounds from the original published article to fill this information gap in the ChEMBL database.[8,9] However, although allosteric and orthosteric annotations were retrieved, the added value of this information in binding predictions was not assessed. We observed that allosteric compounds bind with a lower absolute affinity on average, but with similar ligand efficiency, and we hypothesized that inclusion of this information would lead to a better prediction of affinity.[8]

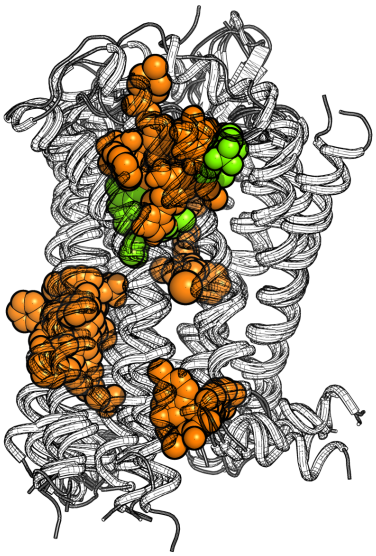


Figure 6.1. Ligand binding sites observed in crystal structures of class A GPCRs. Orthosteric ligands are shown in green, allosteric modulators in orange. The following crystal structures are included (PDB): 4MBS[10], 4MQT[10], 4N6H[11], 4NTJ[12], 4PHU[13], 5LWE[14], 5NDZ[15], 5NLX[16], 5T1A[2], 5TZR[17], 5TZY[17], 5X7D[18], 6C1Q[19], and 6C1R[20].

In the current study an updated text mining protocol is presented that annotates orthosteric and allosteric compounds more accurately. Moreover, the dataset is additionally curated for class A GPCRs, improving the dataset's quality for this target type. The value of allosteric annotation of compounds was assessed by training pChEMBL prediction models for class A GPCRs with and without binding type information. Furthermore, models were constructed that were able to predict allosterism of class A GPCR compounds. Additionally, the reader is able to apply text mining on new documents to annotate the binding type of compounds by using the text mining method and keywords that are provided.

Results

Construction of the orthosteric/allosteric binding type dataset using text mining

Keyword-enabled text mining was applied on documents retrieved from ChEMBL[7] to determine the binding type of compounds (see Methods for details). Only the compounds from the assays that were annotated as "primary assay" were categorized as allosteric, orthosteric, bi-topic, or covalent. The primary assay was assigned based on the highest number of compounds per assay of each document. In contrast to previous work where an effort was made to characterize allosteric compounds,[8,9] the bi-topic and covalent classes were added to remove these compounds from the allosteric and orthosteric classes. Upon manual assessment of the text-mined orthosteric and allosteric compounds for class A GPCRs it was observed that one document contained a primary assay with computationally derived activity values instead of experimental binding activities.[21] Therefore, compounds from the secondary assay were selected for this document. Furthermore, 46 histamine H4 compounds that were categorized as allosteric were in fact orthosteric, which was detected when checked manually. These compounds were corrected accordingly. To expand the dataset, additional 'undetermined' class A GPCR documents were manually categorized, which resulted in a further addition of 306 orthosteric and 24 allosteric compounds. For the majority the binding types were well defined, especially for class A GPCRs, as confirmed by manual random sample assessment and positive model outcomes. The total binding type dataset (for all text-mined proteins) contained 201,721 data points (9,390 allosteric, 14,265 orthosteric, and 178,066 undetermined), of which 60,116 defined data points for class A GPCRs (1,442 allosteric, 9,914 orthosteric, and 48,760 undetermined).

Chemical properties of allosteric modulators

The differences between orthosteric and allosteric binders were explored by evaluating the following properties: molecular weight, ALogP, number of hydrogen bond donors and acceptors, activity (pChEMBL), number of rotatable bonds, and number of aromatic bonds. It was observed that the properties of the allosteric modulators for all proteins in ChEMBL were comparable to those of the orthosteric compounds (**Table 6.1**). In general, the orthosteric compounds covered a wider range of all properties: additional higher molecular weights, more diversity in ALogP, and compounds with more hydrogen bond donors and acceptors. A slight difference was observed in the spread of bioactivities of the two binding types. Although the orthosteric compounds included a higher number of low pChEMBL values (pChEMBL < 4), the bioactivity median was 0.7 log units higher than that of allosteric compounds (pChEMBL 7.1 and 6.4, respectively), in line with previous work.[8] Moreover it was observed that allosteric modulators contained a bigger fraction of aromatic bonds (value normalized to the number of

bonds) compared to orthosteric compounds. This corresponds with previous statements that allosteric modulators are more aromatic.[8,9] Nonetheless, the physicochemical properties of the allosteric modulators are generally not very different from those of orthosteric compounds.

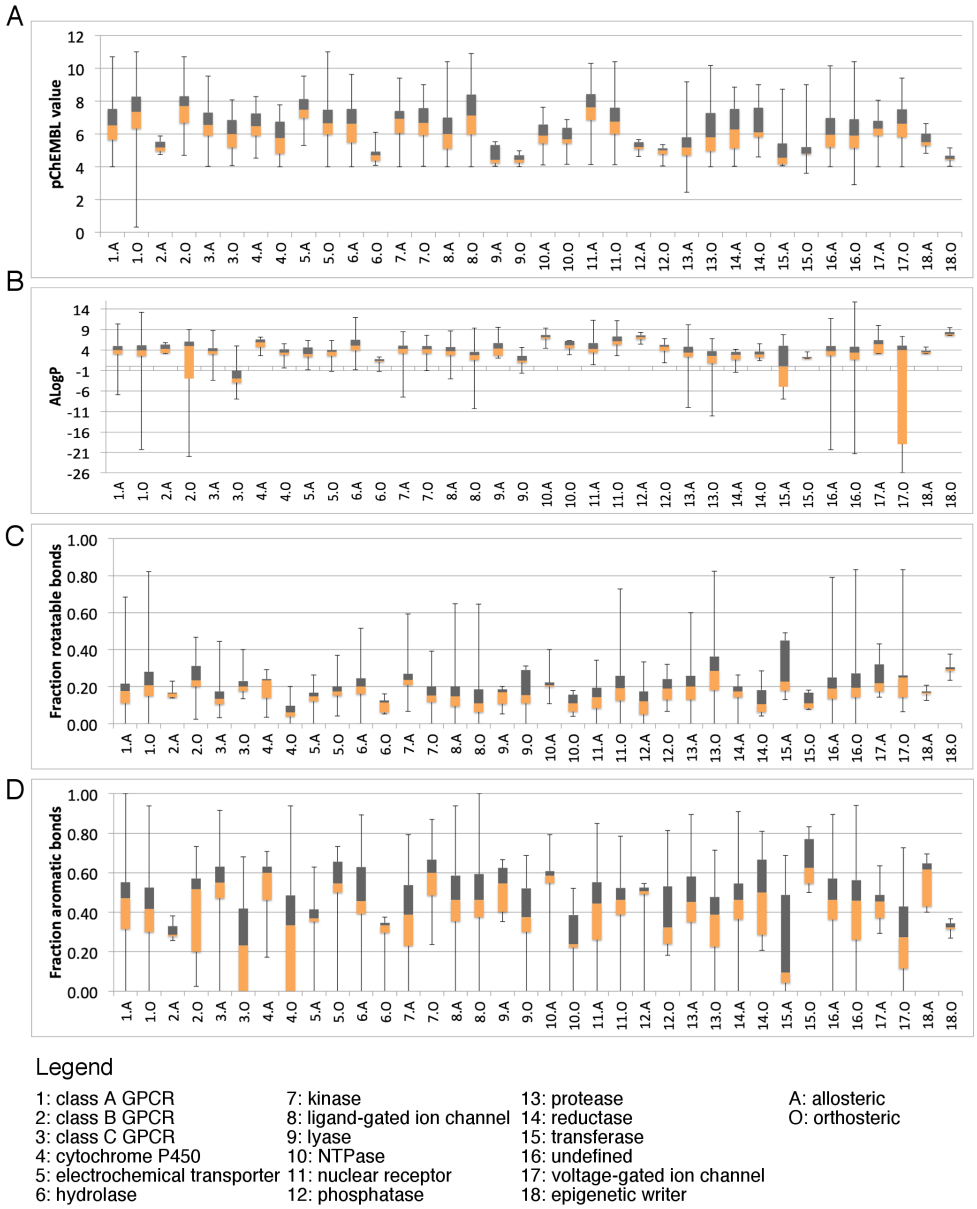
Table 6.1. Physicochemical properties of orthosteric and allosteric compounds in ChEMBL.

	Allosteric			Orthosteric		
	Median	MAD	Mean and standard deviation	Median	MAD	Mean and standard deviation
pChEMBL	6.4	0.9	6.5 ±1.2	7.1	1.0	7.1 ±1.4
Molecular weight	383	61	404 ±135	434	93	541 ±474
ALogP	3.8	1.0	3.8 ±2.0	3.8	1.3	3.1 ±3.9
Num. H donors	1	1	1 ±2	1	1	4 ±8
Num. H acceptors	4	1	4 ±2	5	2	6 ±8
Fraction rotatable bonds	0.17	0.05	0.18 ±0.09	0.20	0.06	0.22 ±0.12
Fraction aromatic bonds	0.50	0.11	0.47 ±0.17	0.43	0.12	0.42 ±0.18

The fraction of rotatable bonds and aromatic bonds represent the number of bonds normalized to the total number of bonds per compound. MAD = mean absolute deviation.

When only a specific protein class was explored, class A GPCRs in this case, it was observed that the properties of the orthosteric and allosteric compounds were comparable to the mean of compounds for all proteins in ChEMBL. However, we noticed that the properties of individual protein classes could vary significantly compared to the properties of the entire set. **Figure 6.2** (and **Supplementary Figure S6.1**) displays the variances between the ligand properties of different protein classes. Although class A GPCRs contained the second-largest number of mined allosteric modulators (1,442 compounds) and is therefore able to cover more chemical ground, it is remarkable that most protein classes have less pronounced properties: the minimum and maximum values for especially molecular weight, ALogP, and hydrogen bond donors and acceptors were less prominent than for the class A GPCRs. The ALogP values of the orthosteric class B and C GPCR compounds (ALogP 5.2 and ALogP -2.7, respectively) were deviating a lot from the conclusion drawn from the entire set where the median ALogP was 3.8. However, these observations could be explained, as shown below.

It was observed that the biggest allosteric set was retrieved for class C GPCRs (2,513 compounds), of which the orthosteric site is located in the extracellular Venus flytrap domain, in contrast to other GPCR classes. The orthosteric compounds in this class had a observably lower ALogP than the allosteric modulators: ALogP median orthosteric -2.7 and allosteric 3.7. Given that the orthosteric compounds for class C GPCRs bind to a structurally different binding site than the allosteric modulators for this receptor, this observation is not surprising. Furthermore, it is known that glutamate receptors make up a large fraction of class C GPCRs, for which the main natural ligand is glutamate, a charged amino acid.[22] Hence, orthosteric ligands should resemble these characteristics, leading to the observed shift in ALogP values (the median ALogP of orthosteric compounds binding to glutamate receptors is -3.0). Moreover, it is noteworthy to mention that the class C GPCR compounds might bias the full dataset as they represent 27%



Legend

1: class A GPCR
2: class B GPCR
3: class C GPCR
4: cytochrome P450
5: electrochemical transporter
6: hydrolase
7: kinase
8: ligand-gated ion channel
9: lyase
10: NTPase
11: nuclear receptor
12: phosphatase
13: protease
14: reductase
15: transferase
16: undefined
17: voltage-gated ion channel
18: epigenetic writer

A: allosteric
O: orthosteric

Figure 6.2. Physicochemical properties of orthosteric and allosteric compounds in ChEMBL per protein family. Orange boxes indicate the area between the 25th percentile and 50th percentile, grey indicates the area between the 50th percentile and 75th percentile, the border between orange and grey indicates the median value or 50th percentile, and the whiskers indicate the minimum value up to the 25th percentile (bottom whiskers) and the 75th percentile up to the maximum value (top whiskers). A) pChEMBL value; B) ALogP; C) fraction of rotatable bonds; and D) fraction of aromatic bonds.

of the allosteric modulators. A similar trend was observed for class B GPCRs, with an ALogP median for orthosteric compounds of 5.2. It was found that the orthosteric compounds for class B GPCRs consisted of solely peptides and that the allosteric class was only represented by seven modulators, which were also peptides. Therefore, class B GPCRs alone will not be a good use case for modeling allosterism.

Class A GPCRs on the other hand, contained a sufficient number of orthosteric and allosteric compounds (9,914 orthosteric and 1,442 allosteric), and a well-balanced portion of small molecules and peptides: orthosteric 5,428 small molecules, 4,395 peptides, and 91 undefined; and allosteric 851 small molecules and 591 peptides (as classified by ChEMBL).[23] It should be noted that the results in this section are based on text-mined binding types. Therefore, caution is advised when comparing and drawing conclusions from the differences observed between orthosteric and allosteric compounds, and also between protein classes. Nonetheless, an effort was made to curate the binding types for class A GPCRs with higher accuracy to support the findings in the current study.

Chemical structures of class A GPCR allosteric modulators

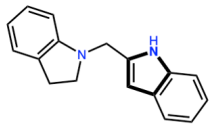
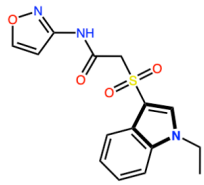
The chemical structures of compounds for the class A GPCRs were explored based on their scaffold trees. The scaffolds were clustered based on two tree levels: the root (smallest scaffold) and the Bemis-Murcko framework[24] (biggest scaffold). Clustering of the class A GPCR set resulted in 313 unique root scaffolds and 4,362 Bemis-Murcko scaffolds. The allosteric modulators covered 19 root scaffolds and 564 Bemis-Murcko scaffolds. The orthosteric compounds contained significantly more unique scaffolds: 229 root and 3770 Bemis-Murcko scaffolds. Furthermore, it was observed that some clusters contained both orthosteric and allosteric compounds: 65 root and 28 Bemis-murcko scaffold clusters were shared. Notably, clusters formed based on the root scaffold are not uniform with the clusters based on Bemis-Murcko scaffolds. In contrast, it was observed that compounds that were grouped into a shared cluster based on root scaffold, were either clustered into an orthosteric-specific or allosteric-specific cluster based on Bemis-Murcko scaffold (**Figure 6.3**). This can be explained since the Bemis-Murcko scaffolds are bigger and therefore, more specific. The root scaffold, however, is more general and therefore prone to cluster more diverse compounds into the same cluster.

Predicting allosteric modulation

Convolutional deep neural networks (DNNs, see Methods for network architecture) were applied to observe if allosteric modulation of compounds for class A GPCRs could be predicted. Compounds of the binding type dataset that were published before the year 2013 were used for model training. Compounds published in 2013 or later were used for validation. This time-split, or temporal split, approach has advantages over random-split validation as it gives a more accurate representation of the models' predictive performances.[25]

The allosteric and orthosteric binding types were predicted by a model trained on physicochemical properties, compound fingerprints, and protein descriptors (for details see Methods). The temporal split validation showed that the proteochemometric (PCM)[26] DNN model, which was trained on compound and protein class A GPCR data, could be used to differentiate allosteric modulators from orthosteric compounds: Matthews Correlation Coefficient (MCC) = 0.54,

sensitivity = 0.54, specificity = 0.94, Positive Predictive Value (PPV) = 0.82, and Negative Predictive Value (NPV) = 0.80. Although sensitivity for allosteric compounds is significantly lower than that for orthosteric compounds (the specificity metric in this case), the model's predictive ability is sufficient to differentiate between both binding types.

Divergent scaffold clusters		
Scaffold		
Root cluster	Shared	Shared
Bemis-Murcko cluster	Orthosteric	Allosteric
Protein	Melatonin receptor 1A	Muscarinic acetylcholine receptor M1

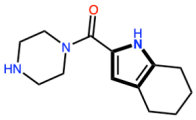
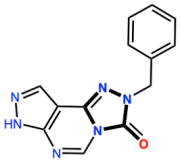
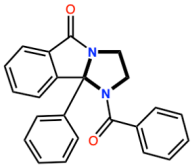
Consistent scaffold clusters			
Scaffold			
Root cluster	Shared	Orthosteric	Allosteric
Bemis-Murcko cluster	Shared	Orthosteric	Allosteric
Protein	Histamine H4 receptor	A_{2A} adenosine receptor	Muscarinic acetylcholine receptor M5

Figure 6.3. Clustered scaffolds of orthosteric and allosteric compounds for class A GPCRs. Bemis-Murcko scaffolds are depicted with the root scaffold indicated in bold.

DNNs outperform Random Forest in binding type predictions

The binding type DNN model was compared to a Random Forest (RF) model that was trained and tested on exactly the same data (class A GPCRs). The predictive performance of the RF model for allosteric compounds was slightly worse than that of the DNN model with a performance of MCC = 0.43, sensitivity = 0.35, specificity = 0.97, PPV = 0.85, and NPV = 0.74. This result indicates that the DNN model was better than the RF model in identification of the binding type of class A GPCRs.

Enhancement of bioactivity predictions by incorporation of binding types

It was studied if the text mining derived binding types could contribute to the predictive performance of pChEMBL prediction models. Instead of the prediction of binding types, the DNN models were trained to predict the activity of compounds (pChEMBL value ≥ 6.5 is

active), using the binding types as descriptor. The models were validated using the same temporal split as mentioned previously (split on the year 2013). The effect of implementation of binding types was analyzed by comparison of the performances of the pChEMBL prediction model with and without binding type. The models were trained on both orthosteric and allosteric compounds, but validated on allosteric modulators and orthosteric compounds separately to remove any bias that might result from the addition of the binding type descriptors. Potential bias might be induced since the allosteric compounds are on average less potent than the orthosteric compounds: pChEMBL 6.6 ±1.3, and 7.3 ±1.3, respectively for class A GPCRs. The findings for bioactivity predictions for allosteric modulators are reported in this section and **Table 6.2**, the results for orthosteric compounds are included in **Supplementary Table S6.2**. It was observed that the MCC of the class A GPCR pChEMBL prediction model with implemented binding type descriptors was slightly higher than for the pChEMBL prediction model that lacked these descriptors: MCC 0.27 and 0.30, respectively for allosteric modulators. Remarkably overall performance (MCC) increased slightly for allosteric bioactivity predictions, but sensitivity decreased (-0.05) when adding binding types. Nevertheless, specificity and accuracy both increased: specificity +0.12 and accuracy +0.03. The receiver operating characteristic (ROC) remained comparable (or slightly worse -0.01).

The addition of the binding types and their effect on pChEMBL predictions for allosteric modulators were further evaluated when the model was trained on the full GPCR dataset and on the dataset containing all proteins in ChEMBL. The pChEMBL predictions for class A GPCRs increased when binding types were included in the models trained on these bigger datasets (**Figure 6.4**). All metrics improved or performed equally upon the addition of binding types to these enlarged datasets. The pChEMBL predictions for allosteric modulators of class A GPCRs when trained on the full GPCR dataset improved the most with a MCC of 0.27 increasing to

Table 6.2. DNN regression model performances for prediction of bioactivities of allosteric modulators for class A GPCRs.

Training dataset	Added descriptor	MCC	Sensitivity	Specificity	Accuracy	PPV	NPV	RMSE	ROC
Class A GPCRs	-	0.27	0.81	0.44	0.61	0.55	0.74	1.07	0.73
	binding type	0.30	0.74	0.56	0.64	0.58	0.72	1.15	0.72
	predicted binding type	0.22	0.51	0.70	0.62	0.59	0.63	1.01	0.70
GPCRs	-	0.27	0.78	0.48	0.62	0.56	0.73	1.01	0.74
	binding type	0.34	0.78	0.55	0.65	0.59	0.75	1.09	0.73
	predicted binding type	0.15	0.52	0.63	0.58	0.54	0.61	0.99	0.67
All proteins in ChEMBL	-	0.35	0.74	0.61	0.67	0.61	0.74	1.03	0.74
	binding type	0.37	0.74	0.62	0.68	0.62	0.74	0.98	0.75
	predicted binding type	0.18	0.43	0.74	0.60	0.58	0.61	1.06	0.67

MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, NPV = Negative Predictive Value, RMSE = Root Mean Square Error, and ROC = receiver operating characteristic.

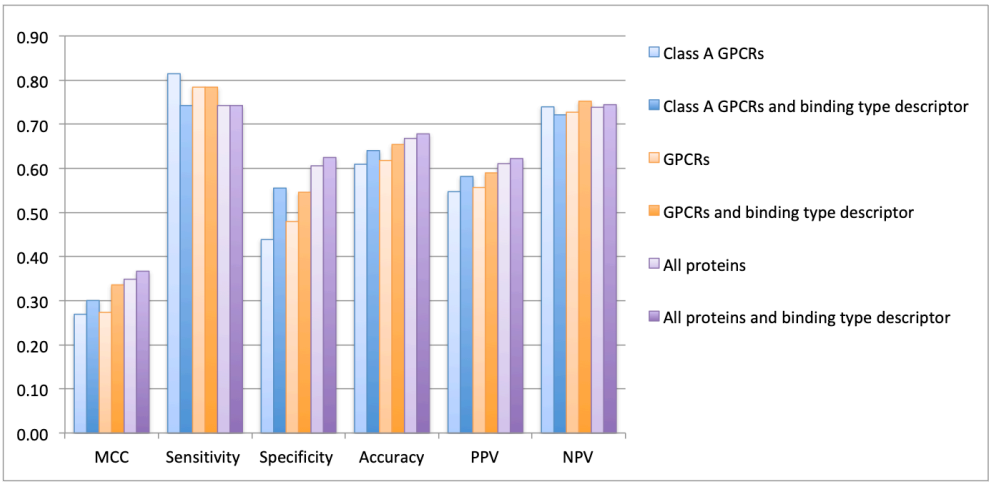


Figure 6.4. The performances of DNN regression models for the prediction of bioactivities of class A GPCR allosteric modulators with and without binding type descriptors. MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, and NPV = Negative Predictive Value.

0.34. The best performing model for allosteric modulators for class A GPCRs was the PCM model trained on all proteins with binding types included (MCC 0.37).

Expansion of binding type predictions into bioactivity models

It was observed that the DNN models were able to predict if compounds were binding in an allosteric or orthosteric manner. Furthermore, the addition of binding type descriptors into pChEMBL prediction models enhanced model performance. Therefore, models were applied iteratively by first predicting the binding types of compounds, subsequently followed by implementation of these predictions into training of a pChEMBL prediction model. The compounds used for testing were retrieved from ChEMBL assays that were categorized as 'undetermined' by the text mining protocol. The predicted binding types were added as descriptor which ranged from zero ("0") to one ("1"), with "1" being allosteric and "0" being orthosteric. The scores for this descriptor were derived as output from the binding type classification model and indicated the probability of a compound belonging to the orthosteric class (score < 0.5) and allosteric class (score > 0.5). The trained class A GPCR pChEMBL prediction model (regression) was compared to a pChEMBL prediction model that lacked these predicted binding type descriptors. The performance was MCC 0.22, sensitivity 0.51, specificity 0.70, accuracy 0.62, PPV 0.59, NPV 0.63, root mean square error 1.01 (RMSE), and ROC 0.70 (Table 6.2). The pChEMBL prediction model that included predicted binding types overall performed worse than the pChEMBL prediction model without the predicted binding types, with the exception of increased specificity and PPV. This may be a result of the implementation of "uncertain" values to generate new predictions; an additional modeling error is introduced into the pChEMBL prediction model that uses predicted descriptors. Furthermore, it was observed that the performances of the models that included predicted binding types decreased when the training set was enlarged with all GPCRs and all proteins in ChEMBL. This is in contrast with the previous observation that model performance increased upon training set enlargement. This indicates that the binding type descriptor weighs heavily in the determination of the pChEMBL

value. Therefore, it is of importance that accurate experimentally validated binding types are used in model training, as predicted binding types are not sufficient.

Discussion

Via our text mining protocol we were able to retrieve accurate "orthosteric" and "allosteric" binding types. The addition of these literature mined binding types in pChEMBL prediction models increased model performance. However, an important aspect is that the binding types should be reliable. Implementation of predicted binding types as descriptor, with a potential error in prediction, did not improve model performance. Although the binding type model did not enhance the accuracy of pChEMBL predictions, the model could be applied to indicate possible allosterism of compounds, as the performance of the binding type model (MCC = 0.54) is remarkably better than random (MCC = 0). In contrast to our study, allosterism-predicting models that are reported in literature are not specified to class A GPCRs or report a less reliable validation method (random split instead of temporal split).[8,27] Therefore, the performance of these models cannot be compared directly.

The physicochemical properties of allosteric modulators generally match the properties of orthosteric compounds, both are drug like.[8] Nevertheless, the chemical features can be distinctive between protein classes, which is reflected by the differences in ALogP, fraction of rotatable bonds, and fraction of aromatic bonds.[9] For class A GPCRs, root and Bemis-Murcko scaffolds were identified that specifically match allosteric modulators or orthosteric compounds. Additionally, scaffolds were found that were shared among allosteric and orthosteric compounds. Manual evaluation of the text mining results showed that some documents were misclassified. This indicates that the text mining derived binding types are not 100% accurate and therefore might still contain some inaccuracies. However, the majority of the binding types are well defined, especially for class A GPCRs, as confirmed by manual random sample assessment and positive model outcomes. Furthermore, the addition of specific keywords to filter for covalent and bi-topic compounds increases the accuracy of categorization of the allosteric and orthosteric classes. Although addition of a bi-topic class in text mining for allosteric modulators has been reported before,[8] implementation of a filter for covalent compounds is only presented in this study. The effort that was made to text mine for binding types has the advantage that multiple binding types can be assembled and annotated. In contrast to the compounds published in the Allosteric Database[28], which only lists allosteric molecules, orthosteric compounds were also actively annotated in this study and thus available as an additional subset.

The dataset that resulted from text mining was not limited to class A GPCRs, but contained determined binding types for multiple protein classes in ChEMBL including receptors, enzymes, and solute carriers. Furthermore, some compounds could not be divided into one of the assigned binding type classes, resulting in 'undetermined binding type' compounds. These compounds were thus not picked up using the extensive list of keywords. Expansion of the searched text from title and abstract text to full paper text may decrease the number of undetermined compounds. However, feasibility may be difficult as not all papers are open access.

It was shown that expansion of the training set using all proteins increased pChEMBL predictions for class A GPCRs, a finding that has been observed for PCM modeling previously.[26] Moreover,

implementation of the text mining derived binding types increased model performance even further. Previous research shows that the choice of descriptors can influence model performance.[29] Nevertheless, to the best of the authors' knowledge, the use of binding types as additional descriptor in bioactivity modeling has not yet been reported.

Conclusions

We successfully retrieved allosteric and orthosteric binding types using text mining. Implementation of these binding types into class A GPCR pChEMBL prediction models improved model performance. Although we were able to predict the binding type of compounds, implementation of these predicted binding types into pChEMBL prediction models did not enhance model performance. Since we did observe improvement of model performance when accurate binding types were included, we encourage the use of binding type descriptors in computational models. Unfortunately, binding type information of compounds is not always readily accessible, especially in the public domain. This research therefore stresses the importance of binding type integration in public databases.

Methods

Text mining and compound classification

Binding types, orthosteric or allosteric, were assigned to the compounds using text mining. Documents that contained both a title and abstract were retrieved from the ChEMBL database (version 22)[23,30,31]. The abstracts and titles were screened for keywords to classify the corresponding document in one of the following classes: orthosteric, allosteric, bi-topic, and covalent. Documents were classified using a 3-layer approach. Remaining, or unclassified documents, were termed "undetermined". Every layer of the classification process included a different set of keywords (**Supplementary Table S6.3**), with the most accurate and general keywords in the first layer, and decreased accuracy of keywords in the following layer. The third layer contained general orthosteric keywords that were not applied in the first layer as they were generally only mentioned in conjunction with allosteric keywords. The documents that could not be classified within the first layer, continued to the second layer, and if applicable, to the third layer. Every document that contained at least one bi-topic or covalent keyword was directly classified into the corresponding class. Classification of orthosteric and allosteric documents was determined by majority vote based on the number of occurrences of keywords belonging to the orthosteric or allosteric class. Compounds from the primary assay, the assay containing the most bioactivities, of the orthosteric and allosteric classified documents were retrieved from ChEMBL and assigned to the same class as the corresponding document. Additionally, class A GPCR compounds from the primary assays of previously undetermined documents were placed in the orthosteric class if published before 2000, as from this year on the literature started more actively reporting allosteric binding types. Therefore, we assumed that the study was performed for orthosteric binders, when allosterism was not explicitly mentioned in a paper in the years prior to 2000.

Dataset

The ChEMBL[7] data derived through text mining was filtered on confidence score 5 (multiple direct proteins), 7 (direct protein complex subunits), and 9 (direct single protein). Additionally, values

retrieved from PubChem[32] and compounds without a defined or exact pChEMBL value (relation is "=") were discarded. Only pChEMBL values that were determined using binding assays were kept. Mean pChEMBL values were calculated for bioactivities of duplicate compounds tested on the same target. All compounds were standardized using Pipeline Pilot (version 16.2.0.58)[33]: the largest fragment was kept, and stereochemistry and charges were standardized.[i]

Compound descriptors

Compounds were encoded by their physicochemical properties.[ii] Additionally FCFP6 fingerprints[34] were used to describe the compounds' structures. Moreover, the activity types (K_i, K_d, EC_{50}, IC_{50} and AC_{50}) were added as binary descriptor, with bit "1" if this activity type was used for measuring the activity value of the compound. The non-binary properties were scaled to values between 0 and 1, where 1 is equal to the highest value obtained for that particular property from ChEMBL.

Protein descriptors

The protein sequences were transformed into descriptors by dividing every sequence into sections that contained five amino acids per section. From these sections the mean values were calculated for the amino acids' physico-chemical properties: molecular weight, number of chiral atoms, LogD, charge, number of hydrogen bonds and acceptors, rigidity, and number of aromatic bonds. Additionally, protein classification levels L1, L2, L3, L4, L5, and L6 from ChEMBL[23] were translated into binary values and added as protein descriptor.

Model training

The RF models were trained using the ensemble package from the python Scikit-Learn module.[35] The amount of trees was set to 500 with random split using seed 12345. Convolutional DNN models were trained to predict binding type and pChEMBL values. The multitask DNN models contained two pyramidal shared hidden layers with 2000 and 100 units respectively. The DNN implementation was built using python 2.7 on Keras with Theano[36] backend. The following settings were applied: test-size = 0.01 (10%), layer dropout = 0.1, learning rate = 0.0001, momentum = 0.8, patience = 25 seconds, maximum epochs = 2000, and maximum batch size = 128. The training set was balanced for the binding type classes (orthosteric and allosteric) using oversampling of the underrepresented class.

Model validation

Models were validated with a temporal split approach. The training set consisted of all data from before the year 2013. The test set comprised all data from the publication year 2013 and above. The test set encompassed approximately 15% of the whole class A GPCR dataset. The performance of both classification (binding type models) and regression models (pChEMBL prediction models) was evaluated using the metrics MCC, sensitivity, specificity, accuracy, PPV, and NPV.

---

i The dataset is freely available at doi: 10.4121/uuid:de9e9805-916f-47e5-8e72-323f675d5d5a.

ii Descriptors are accesible at doi: 10.4121/uuid:5738caea-2390-4dc7-9830-8d9644232144.

References

(1) Soudijn, W.; van Wijngaarden, I.; IJzerman, A. P. Allosteric Modulation of G Protein-Coupled Receptors: Perspectives and Recent Developments. Drug Discov. Today 2004, 9, 752–758. https://doi.org/10.1016/S1359-6446(04)03220-9.

(2) Zheng, Y.; Qin, L.; Zacarías, N. V. O.; de Vries, H.; Han, G. W.; Gustavsson, M.; Dabros, M.; Zhao, C.; Cherney, R. J.; Carter, P.; Stamos, D.; Abagyan, R.; Cherezov, V.; Stevens, R. C.; IJzerman, A. P.; Heitman, L. H.; Tebben, A.; Kufareva, I.; Handel, T. M. Structure of CC Chemokine Receptor 2 with Orthosteric and Allosteric Antagonists. Nature 2016, 540, 458–461. https://doi.org/10.1038/nature20605.

(3) Lu, S.; Zhang, J. Small Molecule Allosteric Modulators of G-Protein-Coupled Receptors - Drug-Target Interactions. J. Med. Chem. 2018, 62. https://doi.org/10.1021/acs.jmedchem.7b01844.

(4) Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P. E. The Protein Data Bank www.rcsb.org.

(5) Congreve, M.; Oswald, C.; Marshall, F. H. Applying Structure-Based Drug Design Approaches to Allosteric Modulators of GPCRs. Trends Pharmacol. Sci. 2017, 38, 837–847. https://doi.org/0.1016/j.tips.2017.05.010.

(6) Wu, H.; Wang, C.; Gregory, K. J.; Han, G. W.; Cho, H. P.; Xia, Y.; Niswender, C. M.; Katritch, V.; Meiler, J.; Cherezov, V.; Conn, P. J.; Stevens, R. C. Structure of a Class C GPCR Metabotropic Glutamate Receptor 1 Bound to an Allosteric Modulator. Science (80-.). 2014, 344, 58 LP – 64. https://doi.org/10.1126/science.1249489.

(7) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100-1107.

(8) van Westen, G. J. P.; Gaulton, A.; Overington, J. P. Chemical, Target, and Bioactive Properties of Allosteric Modulation. PLoS Comput. Biol. 2014, 10, e1003559. https://doi.org/10.1371/journal.pcbi.1003559.

(9) Smith, R. D.; Lu, J.; Carlson, H. A. Are There Physicochemical Differences between Allosteric and Competitive Ligands? PLOS Comput. Biol. 2017, 13, e1005813.

(10) Tan, Q.; Zhu, Y.; Li, J.; Chen, Z.; Han, G. W.; Kufareva, I.; Li, T.; Ma, L.; Fenalti, G.; Li, J.; Zhang, W.; Xie, X.; Yang, H.; Jiang, H.; Cherezov, V.; Liu, H.; Stevens, R. C.; Zhao, Q.; Wu, B. Structure of the CCR5 Chemokine Receptor–HIV Entry Inhibitor Maraviroc Complex. Science (80-.). 2013, 341, 1387 LP – 1390. https://doi.org/10.1126/science.1241475.

(11) Fenalti, G.; Giguere, P. M.; Katritch, V.; Huang, X.-P.; Thompson, A. A.; Cherezov, V.; Roth, B. L.; Stevens, R. C. Molecular Control of δ-Opioid Receptor Signalling. Nature 2014, 506, 191–196. https://doi.org/10.1038/nature12944.

(12) Zhang, K.; Zhang, J.; Gao, Z.-G.; Zhang, D.; Zhu, L.; Han, G. W.; Moss, S. M.; Paoletta, S.; Kiselev, E.; Lu, W.; Fenalti, G.; Zhang, W.; Müller, C. E.; Yang, H.; Jiang, H.; Cherezov, V.; Katritch, V.; Jacobson, K. A.; Stevens, R. C.; Wu, B.; Zhao, Q. Structure of the Human P2Y12 Receptor in Complex with an Antithrombotic Drug. Nature 2014, 509, 115–118. https://doi.org/10.1038/nature13083.

(13) Srivastava, A.; Yano, J.; Hirozane, Y.; Kefala, G.; Gruswitz, F.; Snell, G.; Lane, W.; Ivetac, A.; Aertgeerts, K.; Nguyen, J.; Jennings, A.; Okada, K. High-Resolution Structure of the Human GPR40 Receptor Bound to Allosteric Agonist TAK-875. Nature 2014, 513, 124–127. https://doi.org/10.1038/nature13494.

(14) Oswald, C.; Rappas, M.; Kean, J.; Doré, A. S.; Errey, J. C.; Bennett, K.; Deflorian, F.; Christopher, J. A.; Jazayeri, A.; Mason, J. S.; Congreve, M.; Cooke, R. M.; Marshall, F. H. Intracellular Allosteric Antagonism of the CCR9 Receptor. Nature 2016, 540, 462–465. https://doi.org/10.1038/nature20606.

(15) Cheng, R. K. Y.; Fiez-Vandal, C.; Schlenker, O.; Edman, K.; Aggeler, B.; Brown, D. G.; Brown, G. A.; Cooke, R. M.; Dumelin, C. E.; Doré, A. S.; Geschwindner, S.; Grebner, C.; Hermansson, N.-O.; Jazayeri, A.; Johansson, P.; Leong, L.; Prihandoko, R.; Rappas, M.; Soutter, H.; Snijder, A.; Sundström, L.; Tehan, B.; Thornton, P.; Troast, D.; Wiggin, G.; Zhukov, A.; Marshall, F. H.; Dekker, N. Structural Insight into Allosteric Modulation of Protease-Activated Receptor 2. Nature 2017, 545, 112–115. https://doi.org/10.1038/nature22309.

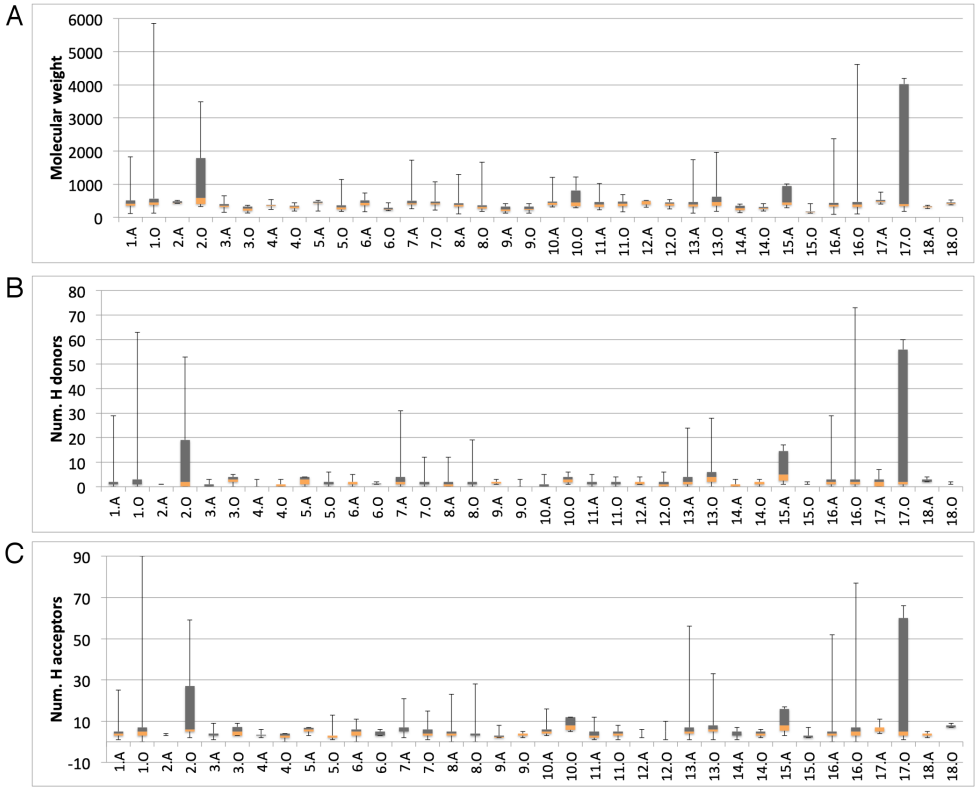
(16) Weinert, T.; Olieric, N.; Cheng, R.; Brünle, S.; James, D.; Ozerov, D.; Gashi, D.; Vera, L.; Marsh, M.; Jaeger, K.; Dworkowski, F.; Panepucci, E.; Basu, S.; Skopintsev, P.; Doré, A. S.; Geng, T.; Cooke, R. M.; Liang, M.; Prota, A. E.; Panneels, V.; Nogly, P.; Ermler, U.; Schertler, G.; Hennig, M.; Steinmetz, M. O.; Wang, M.; Standfuss, J. Serial Millisecond Crystallography for Routine Room-Temperature Structure Determination at Synchrotrons. Nat. Commun. 2017, 8, 542. https://doi.org/10.1038/s41467-017-00630-4.

(17) Lu, J.; Byrne, N.; Wang, J.; Bricogne, G.; Brown, F. K.; Chobanian, H. R.; Colletti, S. L.; Di Salvo, J.; Thomas-Fowlkes, B.; Guo, Y.; Hall, D. L.; Hadix, J.; Hastings, N. B.; Hermes, J. D.; Ho, T.; Howard, A. D.; Josien, H.; Kornienko, M.; Lumb, K. J.; Miller, M. W.; Patel, S. B.; Pio, B.; Plummer, C. W.; Sherborne, B. S.; Sheth, P.; Souza, S.; Tummala, S.; Vonrhein, C.; Webb, M.; Allen, S. J.; Johnston, J. M.; Weinglass, A. B.; Sharma, S.; Soisson, S. M. Structural Basis for the Cooperative Allosteric Activation of the Free Fatty Acid Receptor GPR40. Nat. Struct. Mol. Biol. 2017, 24, 570–577. https://doi.org/10.1038/nsmb.3417.

(18) Liu, X.; Ahn, S.; Kahsai, A. W.; Meng, K.-C.; Latorraca, N. R.; Pani, B.; Venkatakrishnan, A. J.; Masoudi, A.; Weis, W. I.; Dror, R. O.; Chen, X.; Lefkowitz, R. J.; Kobilka, B. K. Mechanism of Intracellular Allosteric B2AR Antagonist Revealed by X-Ray Crystal Structure. Nature 2017, 548, 480–484. https://doi.org/10.1038/nature23652.

(19) Liu, H.; Kim, H. R.; Deepak, R. N. V. K.; Wang, L.; Chung, K. Y.; Fan, H.; Wei, Z.; Zhang, C. Orthosteric and Allosteric Action of the C5a Receptor Antagonists. Nat. Struct. Mol. Biol. 2018, 25, 472–481. https://doi.org/10.1038/s41594-018-0067-z.
(20) Robertson, N.; Rappas, M.; Doré, A. S.; Brown, J.; Bottegoni, G.; Koglin, M.; Cansfield, J.; Jazayeri, A.; Cooke, R. M.; Marshall, F. H. Structure of the Complement C5a Receptor Bound to the Extra-Helical Antagonist NDT9513727. Nature 2018, 553, 111–114. https://doi.org/10.1038/nature25025.
(21) La Regina, G.; Silvestri, R.; Gatti, V.; Lavecchia, A.; Novellino, E.; Befani, O.; Turini, P.; Agostinelli, E. Synthesis, Structure–Activity Relationships and Molecular Modeling Studies of New Indole Inhibitors of Monoamine Oxidases A and B. Bioorg. Med. Chem. 2008, 16, 9729–9740. https://doi.org/10.1016/j.bmc.2008.09.072.
(22) Swanson, G. T.; Sakai, R. Ligands for Ionotropic Glutamate Receptors. Prog. Mol. Subcell. Biol. 2009, 46, 123–157. https://doi.org/10.1007/978-3-540-87895-7_5.
(23) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. Nucleic Acids Res. 2014, 42, D1083-90. https://doi.org/10.1093/nar/gkt1031.
(24) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. 1996, 39, 2887–2893. https://doi.org/10.1021/jm9602928.
(25) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. J. Chem. Inf. Model. 2013, 53, 783–790. https://doi.org/10.1021/ci400084k.
(26) Burggraaff, L.; Oranje, P.; Gouka, R.; van der Pijl, P.; Geldof, M.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 Using Proteochemometric Modeling. J. Cheminform. 2019, 11, 15. https://doi.org/10.1186/s13321-019-0337-8.
(27) Bian, Y.; Jing, Y.; Wang, L.; Ma, S.; Jun, J. J.; Xie, X.-Q. Prediction of Orthosteric and Allosteric Regulations on Cannabinoid Receptors Using Supervised Machine Learning Classifiers. Mol. Pharm. 2019, 16, 2605–2615. https://doi.org/10.1021/acs.molpharmaceut.9b00182.
(28) Shen, Q.; Wang, G.; Li, S.; Liu, X.; Lu, S.; Chen, Z.; Song, K.; Yan, J.; Geng, L.; Huang, Z.; Huang, W.; Chen, G.; Zhang, J. ASD v3.0: Unraveling Allosteric Regulation with Structural Mechanisms and Biological Networks. Nucleic Acids Res. 2015, 44, D527–D535. https://doi.org/10.1093/nar/gkv902.
(29) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. J. Cheminform. 2017, 9, 45. https://doi.org/10.1186/s13321-017-0232-0.
(30) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. Nucleic Acids Res. 2017, 45, D945–D954. https://doi.org/10.1093/nar/gkw1074.
(31) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100–D1107. https://doi.org/10.1093/nar/gkr777.
(32) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. Nucleic Acids Res. 2016, 44, D1202–D1213. https://doi.org/10.1093/nar/gkv951.
(33) BIOVIA Pipeline Pilot. Version 16.2.0.58 http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/.
(34) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742–754. https://doi.org/10.1021/ci100050t.
(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; VanderPlas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. CoRR 2012, abs/1201.0490.
(36) Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.; Ballas, N.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; Bengio, Y.; Bergeron, A.; Zhang, Y.; et al. Theano: A Python Framework for Fast Computation of Mathematical Expressions. 2016.

Supporting Information



Legend

1: class A GPCR	7: kinase	13: protease	A: allosteric
2: class B GPCR	8: ligand-gated ion channel	14: reductase	O: orthosteric
3: class C GPCR	9: lyase	15: transferase	
4: cytochrome P450	10: NTPase	16: undefined	
5: electrochemical	11: nuclear receptor	17: voltage-gated ion channel	
6: hydrolase	12: phosphatase	18: writer	

Figure S6.1. Physicochemical properties of orthosteric and allosteric compounds in ChEMBL per protein family: molecular weight, number of hydrogen bonds and acceptors.

Table S6.2. Performance of bioactivity models for orthosteric compounds.

Training dataset	Added descriptor	MCC	Sensitivity	Specificity	Accuracy	PPV	NPV	RMSE	ROC
Class A GPCRs	-	0.13	0.78	0.34	0.66	0.76	0.37	1.38	0.61
	binding type	0.17	0.82	0.34	0.69	0.77	0.42	1.30	0.63
	predicted binding type	0.16	0.80	0.35	0.68	0.76	0.40	1.32	0.61
GPCRs	-	0.14	0.76	0.37	0.66	0.76	0.38	1.36	0.60
	binding type	0.17	0.83	0.34	0.69	0.77	0.42	1.28	0.63
	predicted binding type	0.09	0.78	0.31	0.65	0.75	0.35	1.27	0.63
All proteins in ChEMBL	-	0.10	0.74	0.35	0.64	0.75	0.35	1.42	0.60
	binding type	0.14	0.74	0.41	0.65	0.77	0.37	1.36	0.61
	predicted binding type	0.14	0.76	0.37	0.65	0.76	0.37	1.31	0.64

MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, NPV = Negative Predictive Value, RMSE = Root Mean Square Error, and ROC = receiver operating characteristic.

Table S6.3. Keywords used in text mining.

Allosteric keywords		
Layer 1	**Layer 2**	**Layer 2**
regulatory site	intracellular domain	Furan-1
secondary binding site	intracellular binding	Brucine
secondary pocket	intracellular ligand	BQCA
nonsubstrate	intracellular site	Tacrine
non-substrate	non-nucleoside analog	Stautosporine
uncompetitive	nonnucleoside analog	Gallamine
un-competitive	NNRTI	Alcuronium
non-competitive	non-nucleoside analog	Heparin
noncompetitive	nonnucleoside analog	Cannabidiol
indirectly inhibit	LUF5484	Loeamide
indirectly activate	PD81723	Antalarmin
indirect activation	Benzodiazepines	L-168049
indirect inhibition	Amiliorides	Fendeline
modulatory actions	Reparixin	Cinacalcet
positive modulator	Agmatine	JNJ16259685
negative modulator	PSNCBAM-1	
receptor modulator	I-TAC	
allosteric modulator	Prichosanthin	
allosteric	Plerixaflor	
allosterism	Trichosanthin	
alosteric	Ancriviroc	
alosterism	Vicriviroc	

allostery	Aplaviroc	
alostery	Maraviroc	

Orthosteric keywords		
Layer 1	**Layer 2**	**Layer 3***
competitive inhibitor	natural ligand	orthostery
competitive activator	primary ligand	orthosterism
competitive inhibition	endogenous ligand	orthosteric
competitive activation	endogenous agonist	orthosterically
competitive-inhibition	endogenous antagonist	
competitive interaction	endogenous blocker	
competitive receptor binding	endogenous activation	
competitive binding	endogenous inhibition	
competitive assay	endogenous binding	
competitive agonist	agonist activity	
competitive antagonist	antagonist activity	
primary binding site	fly trap	
primary binding-site	flytrap	
primary bindingsite	fly-trap	
primary pocket	NRTI	
primary-pocket		
substrate-competitive		
substrate competitive		
directly inhibit		
directly activate		

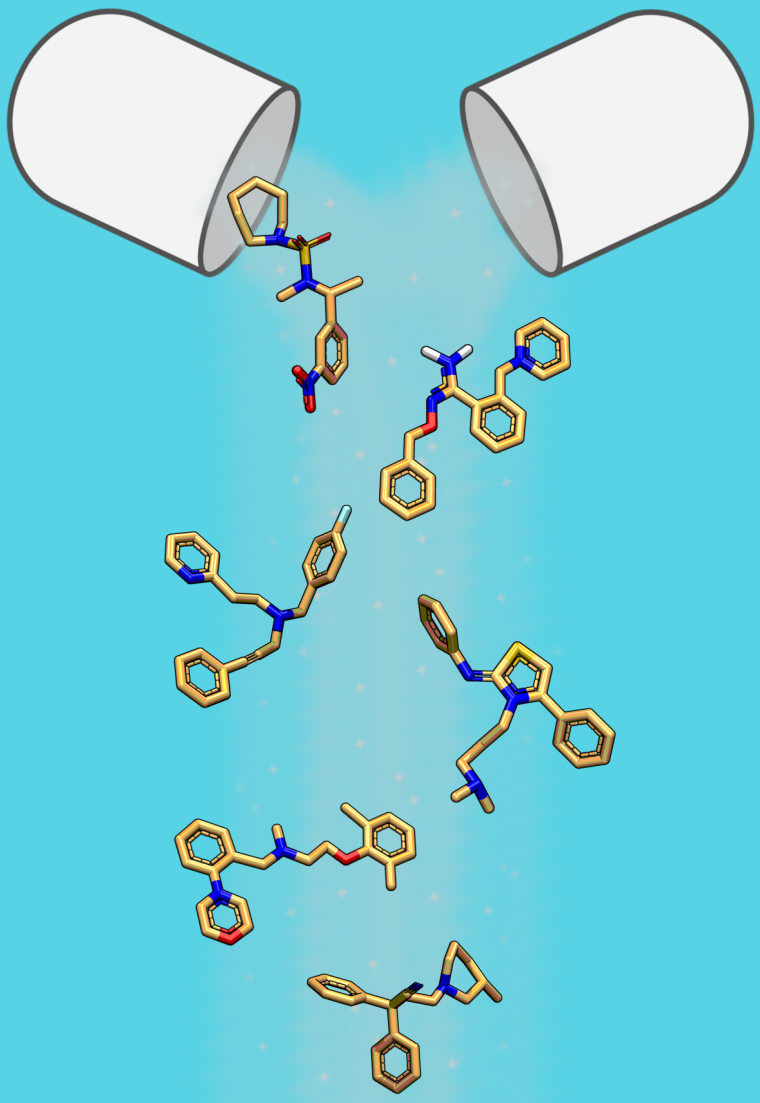
Bi-topic keywords	Covalent keywords	
bisteric	covalent inhibitor	irreversible ligand
dualsteric	covalent inhibition	irreversible target
bivalent	covalent activator	irreversible effect
bitopic	covalent activation	irreversibly
dualsterism	covalent inactivator	warhead
bisterism	covalent inactivation	Neratinib
bitopism	covalent binder	Afatinib
ago-allosteric	covalent binding	BIBW-2992
ago-allosterism	covalent probe	PF-00299804
ago-alosterism	covalent ligand	Carfilzomib
ago-alosteric	covalent target	Abiraterone
ago allosteric	covalent effect	Iniparib
ago allosterism	covalently	BSI201
ago alosteric	irreversible inhibitor	PCI-32765

ago allosterism	irreversible inhibition	Telaprevir
	irreversible activator	Boceprevir
	irreversible activation	Zomepirac
	irreversible inactivator	Daunomycin
	irreversible inactivation	Adriamycin
	irreversible binder	Esomeprazole
	irreversible binding	Clopidogrel
	irreversible probe	Lansoprazole

*Keywords which are (similar to) 'orthosteric' are used in the last layer, when allosteric activities have already been determined

Chapter seven

General Conclusions



Conclusions from this thesis

There is a constant demand for medicines against existing and new diseases.[1–3] Moreover, current drugs can heavily burden patients because of severe side effects.[4] Therefore, we need to develop fast and efficient protocols that can support the discovery of novel drugs. I believe that computational methods can make a substantial contribution to the drug discovery field, as simulations can proceed much faster than real-time experiments and potential hurdles can be foreseen using predictive modeling tools.[5]

One of the major pitfalls in the development of a new drug is the unexpected toxicity of the drug candidate.[6–8] Toxicity is generally manifested as adverse effects, which is caused by off-target binding. These off-targets can be modeled using the computer, which allows us to anticipate on these effects by proposing alterations of the drug candidate. Furthermore, protein target models can be employed to virtually screen for novel compounds that fit the drug-safety requirements (compounds that do not bind to off-targets) and are bioactive on the therapeutic target.[9] This method is not only more time- and cost-efficient than lab experiments, but also has the advantage that it is not limited by the number of compounds that is physically available. This enables us to explore novel chemistry and to identify novel potential drug candidates.

In this thesis, I carried out various modeling approaches to virtually screen for novel compounds. Moreover, I took drug safety into regard by modeling off-targets that can cause adverse effects. The modeling of this so-called selectivity of compounds was explored using case studies on different protein types. The variety of proteins and pathways make that there is no golden standard to obtain selectivity. For example, **Chapter 3** describes how selectivity can be obtained by changing the target site from the kidneys to the small intestine, whereas **Chapter 4** is aimed at achieving selectivity by directly targeting the differences between proteins.

Fast detection of many potential novel selective compounds is feasible by utilization of machine learning as shown in **Chapter 4**. Here, a model is introduced that predicts the selectivity-window between two proteins: the A_1 and A_{2A} adenosine receptors. Although predictive performances were obtained, I envision the model can be optimized further if sufficient compound-protein information is available to implement into the model. The implementation of specific compound-protein information into machine learning models was explored in **Chapter 6**, where bioactivity was modeled and binding types were predicted. Nevertheless, in order to obtain the desired data, an additional step was necessary; by using text mining the binding location of a ligand on a protein was retrieved. This information was included into machine learning models as binding type descriptors and annotated whether a compound binds orthosterically or allosterically. The additional information increased the model performance and thus shows the value of a well-defined dataset and descriptors. Since the binding types are not generally incorporated in accessible biochemical databases[10–12], it is encouraged to include more (meta)data in chemical libraries to make the information easily accessible to all users. This will surely enable the design of high-performance models and therewith, more precise predictions.

The lack of valuable binding type information is only a small fraction of the bigger data issue in computational drug discovery: data is often incomplete and prone to errors as data is collected from different sources and experiments.[13,14] **Chapter 2** highlights this issue and also describes what data is required for the computational modeling techniques that are used throughout

this thesis. Although the methods that are described in **Chapter 2** are generally used to model bioactivity of compounds, this thesis provides the reader with multiple tactics to use these techniques in selectivity modeling. Moreover, the issue of data deficiency is tackled in **Chapter 3**, where compounds were modeled for SGLT1 in the small intestine. Since the available data was insufficient in terms of chemical diversity, additional experiments had to be performed in order to build predictive models.

The importance of having enough data at disposal plays even a larger role in selectivity modeling than in bioactivity modeling. In **Chapter 4** twice as much data was needed to model selectivity compared to bioactivity modeling, as compound data for two proteins is required. Furthermore, **Chapter 5** shows the expansion of a data set that is needed for modeling multi-target drugs. Nevertheless, this thesis provides the reader with handles how data gaps can be filled by additional experiments, or complementary techniques such as proteochemometric modeling.

The vast necessity of data for model building raises the issue of the influence of the used dataset. It is true that one must be aware of prediction bias and the chemical applicability domain of models. This is especially the case for statistical models, which may not be able to correctly identify new entities that are too dissimilar from the initial training data.[15] This issue was touched upon in **Chapter 5**, where statistical models were challenged by the identification of chemically very distinct inhibitors. Here, structure-based models were the key to success in the identification of chemically novel inhibitors.

Based on the challenges that were met during this research it is advised to always carefully evaluate the use of datasets and models of any protein in any type of modeling. The model validations that were performed on all proteins in the rigorous workflow presented in **Chapter 5** made it clear that modeling results vary between proteins. Generally proteins for which less data was available performed worse than proteins that had large datasets. However, results between different models of the same protein also varied greatly in accuracy, an observation that thus is not directly linked to the available data.

Altogether, the findings in this thesis showed me the complexity of computational drug discovery and made it clear that selectivity modeling is a challenge that requires a custom approach for each case separately. Decisions regarding selectivity modeling should be tuned based on *inter alia* the proteins' mechanisms, location, and available data. Although this complex matter still holds a lot of uncovered ground, I strongly believe that in the future we are able to identify and fill the essential data gaps, which will allow us to generate large-scale models that encompass multiple facets of the human physiology. This will reinforce the computational predictions to become an indispensable part of drug discovery.

Future perspectives

Currently the search for new drugs is generally conducted in the range of 10^{33} medicine-like molecules (according to Lipinski's rule of five).[16] However, if computational capacity allows it, it may be worthwhile to search beyond this limit. In the following section I will share my perspective on why this is important and how we can achieve this. Rather than just focusing on selectivity, the points discussed here are relevant to computational drug discovery in general.

By expanding the search range of chemical entities that are considered in virtual screening, very novel drug candidates can be detected. A major benefit of such novel drug candidates is that they can be crucial in diseases that are prone to drug resistance. For instance, the underlying cause of antibiotic resistance is generally the natural modification of bacteria. Bacteria that are adapted are subsequently shielded from being killed by antibiotics. As a result, the used antibiotic is no longer effective against this particular bacteria strain. Important sources of new antibiotics have been natural resources such as plants (horminone) and soil (platensimycin). Such natural sources often present us with antibiotics that are chemically diverse. However, the detection of antibiotics from natural products can be difficult and costly. Therefore, computational screening for chemically novel drugs may be a good addition to provide an outcome for antibiotic resistance.[17]

Moreover, chemically diverse drugs can also be of interest when toxicity needs to be reduced or when selectivity needs to be achieved. Novel chemical compounds can result in entirely different pharmacological profiles, while maintaining a comparable effect on the on-target as existing drugs. By expanding the searched chemical scope, these compounds can be identified.

The simplest and fastest way to upscale compound screening processes is the use of statistical models. Nevertheless, the drawback of this method is the limited chemical space that it can be applied to, since the applicability domain is defined by the data that is used in model training.[15] Therefore, these models may not be very effective in broadening our chemical exploration scope.

Therefore, I propose that the use of structure-based methods is a more fitted approach. Up till now, the bottlenecks in structure-based modeling have been 1) the availability of crystal structures and 2) the speed of the simulations. I foresee that in the fast future both of these issues will no longer be considered as restrictions in structure-based drug discovery. The rise of cryo-electron microscopy (cryo-EM),[18] a technique used to derive protein structures, will aid in the elucidation of many protein structures that are challenging to visualize using other techniques such as X-ray crystallography (XRC) or nuclear magnetic resonance (NMR) microscopy.[19,20] Moreover, cryo-EM accepts lower sample purities and does not rely on the protein being crystalized as it freezes the protein state.[21] Therefore, also larger protein complexes can be visualized using cryo-EM.

Although cryo-EM can aid in solving many protein structures, its usability may be hampered by the lower resolution that it provides compared to XRC. However, my thoughts on this are that cryo-EM and XRC should not be used in parallel, but complementarily.[22] The capability of cryo-EM to capture a wider range of proteins can be empowered by the high resolution of XRC, ultimately revealing more insight into protein structures.

The second bottleneck in structure-based research concerns the limited speed of calculations and simulations. However, the computational capacity has grown significantly over the past years. Moreover, the development of the quantum computer promises great gains in calculation time for structure-based drug discovery.[23] By running multiple calculations in parallel quantum computing is believed to be exponentially faster than traditional hardware.[24]

These innovations will enable us to expand our search and detect potential medicines in a new chemical space. Although structure-based modeling is considered the best way to facilitate this,

machine learning and artificial intelligence will not be disregarded. Statistical models may not be very useful in the identification of novel chemistry, but can come in very handy if given a different purpose. Machine learning can support the discovery of novel drugs by contributing to the optimization of simulations. For example, structure-based scoring functions can be optimized using machine learning and forcefields can be optimized.[25]

Another option might be to enlarge the applicability domain of machine learning models, which will increase the accuracy of prediction of chemically diverse compounds. However, this still will need information derived from structure-based models. The addition of structure-based descriptors into machine learning models might increase model performance, but I feel that this method will be quite protein-specific. Even if the information derived from the protein structure or ligand-protein interactions can easily be added, such models will probably not outperform structure-based models when applied to a diverse chemical set of molecules, as 'contextual' protein information is lacking. Although both ligand and protein information will be included in the model, unique ligand-protein interactions and the ligand conformation will remain difficult to capture in descriptors.

As illustrated in previous paragraphs there is room for improvement of computational innovations. Significant improvements can be made by increasing the data quality and quantity. However, an important, but often overlooked aspect of the contribution of computers in drug discovery lies in the human applicability of the developed methods. I believe that the use of computational methods can be popularized by providing the chemists, biologists, etc., with easy-to-use tools for which no technical computational experience is required. Such a request requires technical operations; therefore data engineers will probably become valuable assets for computational drug discovery groups.

By enabling easy access to computational techniques through tooling, scientific knowledge can swiftly be shared, expanded, and distributed. This development in culture may not only aid fellow scientists in becoming more computational-driven, but also may enable community platforms. Community effort already showed to be valuable in, for example, the discovery of the protein structure of M-PMV retroviral protease. Framed as a web-based game called Foldit, players were challenged to solve the protein structure of the protease.[26] Without any prior biochemical knowledge needed, the players were able to obtain functional protein models. More recently we have witnessed the willingness of community research during the outbreak of COVID-19. The urgency for a cure and vaccine moved researchers to join efforts by sharing data, databases, and even physical compounds.

Sharing information can impact the discovery of new drugs dramatically since knowledge and skills can be combined and reinforced. Let us hope that this pandemic incites academia, but especially pharmaceutical industry, to transform the act of sharing knowledge from being an exception to being the norm.

Final remarks

This thesis highlights the importance of drug selectivity and displays the benefits of computational models in the prediction of ligand-protein binding. It shows us that selectivity modeling requires a customized approach, which should be fitted to the considered targets and available data. We have seen that in order to tune selectivity correspondingly to the pattern of on and off-targets, extensive preliminary validation should be performed to assess if selectivity modeling is possible at all. Altogether, this thesis leaves the reader with a good understanding of (the challenges in) computational drug discovery and provides insights into selectivity modeling.

References

(1) Atkinson, M. A.; Eisenbarth, G. S.; Michels, A. W. Type 1 Diabetes. Lancet 2014, 383, 69–82. https://doi.org/https://doi.org/10.1016/S0140-6736(13)60591-7.
(2) Bluestone, J. A.; Herold, K.; Eisenbarth, G. Genetics, Pathogenesis and Clinical Interventions in Type 1 Diabetes. Nature 2010, 464, 1293–1300. https://doi.org/10.1038/nature08933.
(3) Trépo, C.; Chan, H. L. Y.; Lok, A. Hepatitis B Virus Infection. Lancet 2014, 384, 2053–2063. https://doi.org/https://doi.org/10.1016/S0140-6736(14)60220-8.
(4) Kroschinsky, F.; Stölzel, F.; von Bonin, S.; Beutel, G.; Kochanek, M.; Kiehl, M.; Schellongowski, P.; Group, on behalf of the I. C. in H. and O. P. (iCHOP) C. New Drugs, New Toxicities: Severe Side Effects of Modern Targeted and Immunotherapy of Cancer and Their Management. Crit. Care 2017, 21, 89. https://doi.org/10.1186/s13054-017-1678-1.
(5) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-Driven Medicinal Chemistry in the Era of Big Data. Drug Discov. Today 2014, 19, 859–868. https://doi.org/https://doi.org/10.1016/j.drudis.2013.12.004.
(6) Van Norman, G. A. Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. JACC Basic to Transl. Sci. 2016, 1, 170–179. https://doi.org/https://doi.org/10.1016/j.jacbts.2016.03.002.
(7) Jüni, P.; Nartey, L.; Reichenbach, S.; Sterchi, R.; Dieppe, P. A.; Egger, M. Risk of Cardiovascular Events and Rofecoxib: Cumulative Meta-Analysis. Lancet 2004, 364, 2021–2029. https://doi.org/https://doi.org/10.1016/S0140-6736(04)17514-4.
(8) Siramshetty, V. B.; Nickel, J.; Omieczynski, C.; Gohlke, B.-O.; Drwal, M. N.; Preissner, R. WITHDRAWN—a Resource for Withdrawn and Discontinued Drugs. Nucleic Acids Res. 2015, 44, D1080–D1086. https://doi.org/10.1093/nar/gkv1192.
(9) Lee, H.-M.; Yu, M.-S.; Kazmi, S. R.; Oh, S. Y.; Rhee, K.-H.; Bae, M.-A.; Lee, B. H.; Shin, D.-S.; Oh, K.-S.; Ceong, H.; Lee, D.; Na, D. Computational Determination of HERG-Related Cardiotoxicity of Drug Candidates. BMC Bioinformatics 2019, 20, 250. https://doi.org/10.1186/s12859-019-2814-5.
(10) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. J. Cheminform. 2017, 9, 17. https://doi.org/10.1186/s13321-017-0203-5.
(11) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. Nucleic Acids Res. 2017, 45, D945–D954. https://doi.org/10.1093/nar/gkw1074.
(12) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. Nucleic Acids Res. 2007, 35, D198–D201. https://doi.org/10.1093/nar/gkl999.
(13) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. J. Chem. Inf. Model. 2013, 53, 2499–2505. https://doi.org/10.1021/ci400099q.
(14) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. PLoS One 2013, 8, e61007. https://doi.org/10.1371/journal.pone.0061007.
(15) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. J. Chem. Inf. Model. 2005, 45, 839–849. https://doi.org/10.1021/ci0500381.
(16) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 2001, 46, 3–26.
(17) Wright, G. D.; Sutherland, A. D. New Strategies for Combating Multidrug-Resistant Bacteria. Trends Mol. Med. 2007, 13, 260–267. https://doi.org/https://doi.org/10.1016/j.molmed.2007.04.004.
(18) Fernandez-Leiro, R.; Scheres, S. H. W. Unravelling Biological Macromolecules with Cryo-Electron Microscopy. Nature 2016, 537, 339–346. https://doi.org/10.1038/nature19948.
(19) Burley, S. K. An Overview of Structural Genomics. Nat. Struct. Biol. 2000, 7, 932–934. https://doi.org/10.1038/80697.
(20) Shoemaker, S. C.; Ando, N. X-Rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. Biochemistry 2018, 57, 277–285. https://doi.org/10.1021/acs.biochem.7b01031.
(21) Scapin, G.; Potter, C. S.; Carragher, B. Cryo-EM for Small Molecules Discovery, Design, Understanding, and Application. Cell Chem. Biol. 2018, 25, 1318–1325. https://doi.org/10.1016/j.chembiol.2018.07.006.
(22) Rossmann, M. G.; Morais, M. C.; Leiman, P. G.; Zhang, W. Combining X-Ray Crystallography and Electron Microscopy. Structure 2005, 13, 355–362. https://doi.org/https://doi.org/10.1016/j.str.2005.01.005.
(23) Lanyon, B. P.; Whitfield, J. D.; Gillett, G. G.; Goggin, M. E.; Almeida, M. P.; Kassal, I.; Biamonte, J. D.; Mohseni, M.; Powell, B. J.; Barbieri, M.; Aspuru-Guzik, A.; White, A. G. Towards Quantum Chemistry on a Quantum Computer. Nat.

Chem. 2010, 2, 106–111. https://doi.org/10.1038/nchem.483.
(24) Denchev, V.; Boixo, S.; Isakov, S.; Ding, N.; Babbush, R.; Smelyanskiy, V.; Martinis, J.; Neven, H. What Is the Computational Value of Finite Range Tunneling? Phys. Rev. X 2015, 6. https://doi.org/10.1103/PhysRevX.6.031015.
(25) Batool, M.; Ahmad, B.; Choi, S. A Structure-Based Drug Discovery Paradigm. Int. J. Mol. Sci. 2019, 20, 2783. https://doi.org/10.3390/ijms20112783.
(26) Khatib, F.; DiMaio, F.; Cooper, S.; Kazmierczyk, M.; Gilski, M.; Krzywda, S.; Zabranska, H.; Pichova, I.; Thompson, J.; Popović, Z.; Jaskolski, M.; Baker, D.; Group, F. C.; Group, F. V. C. Crystal Structure of a Monomeric Retroviral Protease Solved by Protein Folding Game Players. Nat. Struct. Mol. Biol. 2011, 18, 1175–1177. https://doi.org/10.1038/nsmb.2119.

Summary

Summary

Computational modeling plays an important role in early drug discovery. Using computers, time spend on medicine development can be decreased significantly. Moreover, since fewer resources are needed to detect potential new medicines, the drug discovery process becomes more cost-effective. This thesis describes how computational models can be applied to enhance the drug discovery process by the prediction of selectivity of compounds, and therewith implicitly toxicity.

The subject of computational medicine development and selectivity is touched upon in **Chapter 1**. Here, general drug discovery concepts are outlined and the role of computers in medicine development is highlighted. Moreover, it is made clear why one would want to obtain selectivity.

Chapter 2 goes into depth on computational methods that can be used in drug discovery. The roles of both ligand- and protein-based methods are discussed in the light of their applicability in target prediction studies. Furthermore, the methods' requirements and restrictions are discussed, which gives the reader a good understanding of the applicability of these tools in drug discovery in general.

A selection of the computational tools that are described in **Chapter 2** are applied in case studies in the subsequent **Chapters 3**, **4**, and **5**. **Chapter 3** is limited to the use of ligand-based methods due to the lack of structure-based protein information. In contrast, **Chapter 4** and **5** describe the use of both ligand- and structure-based methods. With the machine learning method applied in **Chapter 3**, chemically novel SGLT1 inhibitors could be detected. The key to success in this chapter was to create a chemically diverse training set in order to expand the applicability domain.

Ligand-based machine learning was also put to the test in **Chapter 4**, where it was used to directly model the difference of selectivity between two proteins – the A_1 and A_{2A} adenosine receptors. Where traditionally machine learning in drug discovery is used to model bioactivity (as in **Chapter 3**), it was now applied to predict the bioactivity difference. Since this method only scales new compounds based on selectivity, inactive compounds cannot be detected with this model. Therefore, it was combined with structure-based docking models to filter for inactive compounds.

In **Chapter 5**, a different combination of ligand-based and structure-based techniques was applied. A subsequent workflow was developed that utilized machine learning, structure-based docking, and binding pose metadynamics, to detect kinase inhibitors. Here, multiple machine learning models were designed to filter and score compounds. In contrast to **Chapter 4** where docking was used to exclude inactive compounds, in this workflow docking was used to rank compounds. Together with pose metadynamics, a more accurate binding affinity score could be obtained than with docking by itself. The final combined ligand- and structure-based score was used to select novel RET kinase inhibitors.

Chapter 6 addresses a subject that is worth considering when modeling selectivity: the binding location of the ligand on the protein. Some proteins contain multiple binding sites, which can bind chemically different compounds. This can influence the predictive capability of models when not considered. **Chapter 6** displays the effect on model predictive performance when

binding locations on proteins are included in the model. Moreover, it shows that the binding location can be predicted when training data is sufficient.

The final chapter, **Chapter 7**, lists the most important conclusions that can be drawn from this thesis. Attention is given to the applicability of selectivity modeling in different situations. Moreover, both the limitations and opportunities of computational models in a broader perspective are described with an emphasis on available data in terms of quantity and quality (diversity).

Samenvatting

Samenvatting

De bijdrage van computers in de medicijnontwikkeling wordt steeds belangrijker naarmate de computer capaciteit verhoogt en de vraag naar nieuwe en betere medicijnen groter wordt. Door gebruik te maken van computers kan de tijd dat het kost om een medicijn te ontwikkelen aanzienlijk ingekort worden. Bovendien werkt het inzetten van computers en rekenmodellen kostenbesparend omdat er minder uitgegeven wordt aan labmiddelen en mankracht. Dit proefschrift beschrijft hoe computermodellen gebruikt kunnen worden in het medicijnontwikkelingsproces. Hierbij wordt specifiek ingegaan op het modeleren van de selectiviteit van potentiële medicijnen. Deze selectiviteit geeft tevens een indicatie van de toxiciteit van de desbetreffende kandidaat-medicijnen.

Het concept betreffende selectiviteit in de medicijnontwikkeling is toegelicht in **Hoofdstuk 1**. In dit hoofdstuk worden algemene begrippen in de medicijnontwikkeling nader uitgelegd en wordt tevens het gebruik van computers hierin toegelicht. Bovendien wordt het belang van het verkrijgen van selectiviteit aangekaart.

Hoofdstuk 2 gaat in op computertechnieken die kunnen worden toegepast in het ontwikkelen van nieuwe medicijnen waarbij een onderscheid wordt gemaakt tussen modellen gebaseerd op 'compounds' en modellen gebaseerd op eiwitstructuren. De toepassing van beide soorten modellen wordt beschreven in het kader van het voorspellen van potentiële 'target' eiwitten – eiwitten waar kandidaat-medicijnen aan zouden kunnen binden. Verder worden de vereisten en beperkingen van de technieken toegelicht waardoor de lezer inzicht krijgt in de toepasbaarheid van de verschillende modellen.

Een aantal technieken uit **Hoofdstuk 2** worden vervolgens toegepast in **Hoofdstuk 3**, **4**, en **5**. Hoewel **Hoofdstuk 3** is gelimiteerd tot het gebruik van statistische modellen waarbij enkel gebruikt wordt gemaakt van 'compound' informatie, beschrijven **Hoofdstuk 4** en **5** beide een casus waarin zowel statische modellen als eiwitstructuur-gebaseerde modellen worden toegepast.

Door in **Hoofdstuk 3** gebruik te maken van 'machine learning' konden er nieuwe remmers gevonden worden voor het transporteiwit SGLT1, een interessant eiwit in de behandeling van diabetes type II. De sleutel tot succes in dit hoofdstuk was het aanvullen van de model 'training' dataset met een selectie van zeer diverse compounds. Door diverse compounds toe te voegen kon de toepasbaarheid van het machine learning model worden vergroot, wat resulteerde in chemisch nieuwe SGLT1 remmers.

Machine learning werd ook toegepast in **Hoofdstuk 4**. In tegenstelling tot **Hoofdstuk 3** werd het hier niet gebruikt om de bioactiviteit van compounds te voorspellen, maar werd het ingezet om het verschil in bioactiviteit, en dus de selectiviteit, tussen twee eiwitten te modelleren. Het machine learning model werd succesvol toegepast om de selectiviteit te voorspellen tussen de A_1 en A_{2A} adenosine receptoren. Echter, omdat het selectiviteitsmodel niet in staat is inactieve compounds af te vangen, werden A_1 en A_{2A} adenosine structuurmodellen gebruikt om deze eruit te filteren.

Hoofdstuk 5 gebruikt een alternatieve combinatie van machine learning en eiwitstructuurmodellen. Er werd een workflow ontwikkeld waarbij machine learning, structuurmodellen en bindingspose metadynamica werden ingezet om kinase blokkers te vinden. Verschillende machine learning modellen werden toegepast om enerzijds compounds te filteren en anderzijds om compounds te rangschikken. In tegenstelling tot **Hoofdstuk 4** waarbij structuurmodellen werden gebruikt om compounds te filteren, werden de structuurmodellen in dit geval gebruikt om de compounds te ordenen. Door deze techniek te combineren met bindingpose metadynamica konden de compounds nog nauwkeuriger gerangschikt worden. Tenslotte werd de score van de machine learning voorspelling gecombineerd met de score van de structuurmodellen om nieuwe RET kinase blokkers te selecteren.

Een onderwerp dat belangrijk is om in acht te nemen in het modelleren van selectiviteit is de bindingplaats van een compound op een eiwit. Sommige eiwitten bevatten meerdere bindingsplaatsen waarop vaak chemisch diverse compounds kunnen binden. **Hoofdstuk 6** gaat op dit onderwerp in en laat zien dat voorspellingen nauwkeuriger worden wanneer de bindingplaatsinformatie in de modellen worden meegenomen. Bovendien laat dit hoofdstuk zien dat het mogelijk is om de bindingplaats op een eiwit voor compounds te voorspellen als de training dataset toereikend is.

Het afsluitende hoofdstuk, **Hoofdstuk 7**, beschrijft de belangrijkste conclusies van dit proefschrift. Er wordt aandacht besteed aan de conclusies met betrekking op de toepasbaarheid van selectiviteitsmodellen in verschillende situaties. Verder worden de mogelijkheden en beperkingen van computermodellen in het algemeen toegelicht, waarbij met name de hoeveelheid beschikbare data en datakwaliteit aan bod komen.

Samenvatting voor Leken

Samenvatting voor leken

Dit proefschrift beschrijft hoe computers kunnen worden ingezet in het ontdekken en ontwikkelingen van nieuwe medicijnen. Dat het ontwikkelen van medicijnen van groot belang is moge duidelijk zijn. Er zijn namelijk nog steeds ziektes waarvoor er geen medicijnen beschikbaar zijn of de medicijnen die er zijn leiden tot ernstige bijwerkingen.

Om te kunnen begrijpen hoe een nieuw medicijn ontwikkelt kan worden, moeten we eerst weten hoe een medicijn zijn werking uitoefent (**Hoofdstuk 1**). De potentiële medicijnen die in dit proefschrift aan bod komen vallen onder de groep 'kleine moleculen'. Dit is dezelfde klasse als bijvoorbeeld paracetamol en sildenafil (Viagra), maar ook suikers en aminozuren worden beschouwd als kleine moleculen. Dit soort medicijnen zorgen over het algemeen voor een effect in het lichaam door een interactie aan te gaan met specifieke eiwitten. Het menselijk lichaam bestaat uit heel veel verschillende soorten eiwitten, die allemaal hun eigen functie hebben. Sommige eiwitten zijn in staat signalen door te geven aan de rest van het lichaam, terwijl andere eiwitten belangrijk zijn voor de vertering van voedsel. Je kunt je voorstellen dat bij het ontwikkelen van een nieuw medicijn het dus belangrijk is te weten welk effect je wilt bereiken en welk eiwit je daarvoor moet beïnvloeden. Een pijnstiller richt zich bijvoorbeeld op eiwitten die betrokken zijn bij het doorgeven van een pijnsignaal.

Maar hoe weet een medicijn de juiste eiwitten te vinden? En hoe weet een paracetamol dat jij buikpijn hebt en geen kiespijn? In het kort: dat weet het niet. Wat dat betreft zijn medicijnen een beetje dom. Ze gaan waar ze kunnen gaan en worden uiteindelijk door ons lichaam weer uitgescheiden. Onderweg komen de medicijnen verschillende eiwitten tegen en als ze toevallig een bijpassend eiwit vinden kunnen ze daar een interactie mee aangaan. Of een medicijn-eiwit combinatie een interactie oplevert hangt af van verschillende factoren, maar grofweg kan het worden vergeleken met het passen van een sleutel in een slot. In het geval van de paracetamol past het dus op het pijneiwit – een eiwit dat op meerdere plekken in het lichaam voorkomt. Een paracetamol weet echter niet waar het moet zijn en grote kans dat het dus óók aan de pijneiwitten bindt die te maken hebben met kiespijn in plaats van alleen de pijneiwitten in je buik. Dit merk je alleen niet omdat je op dat moment geen kiespijn ervaart en het pijnstillend effect dus niet opvalt.

Tijdens het medicijnontwikkelingsproces proberen we medicijnen zo goed mogelijk te laten passen op de relevante eiwitten. Dit heet ook wel het selectief maken van een medicijn. Wanneer medicijnen namelijk ook binden aan eiwitten waar ze beter niet aan kunnen binden krijg je bijwerkingen. Het binden van een medicijn aan eiwitten die bijvoorbeeld het hart ontregelen kan dramatische gevolgen hebben.

Met de computer kunnen we tot op zeker hoogte simuleren of een stofje (klein molecuul) past op het relevante eiwit. Bovendien kunnen we ook simuleren of het stofje bindt op eiwitten die bijwerkingen veroorzaken. Die laatste stofjes zijn dus minder geschikt als medicijn. Met de computer kunnen er miljoenen stofjes per dag worden getest, waarvan vervolgens alleen de beste worden gekozen om te dienen als potentieel medicijn.

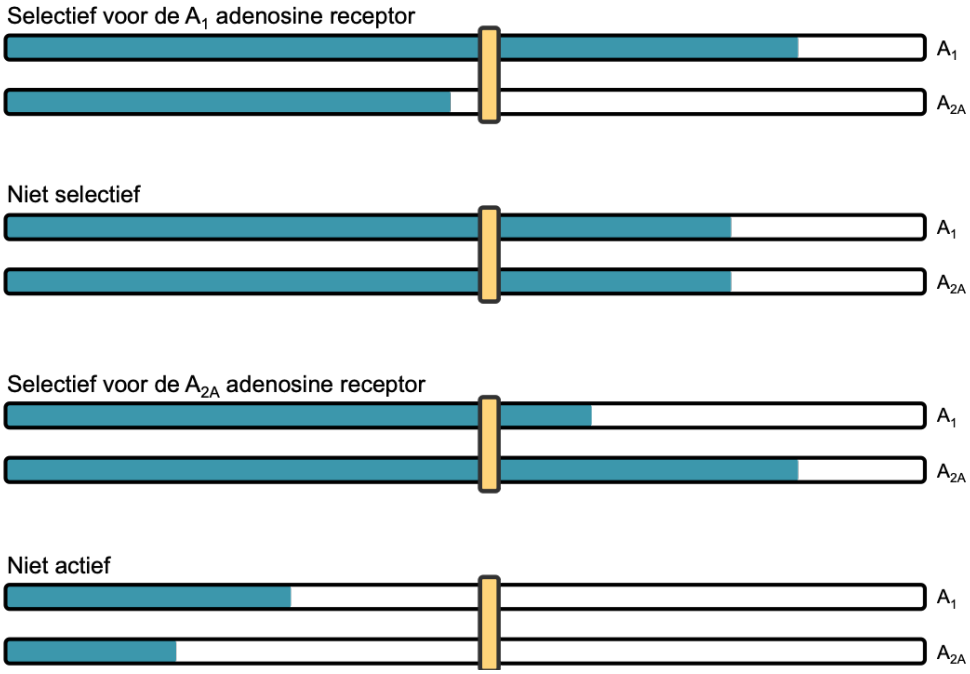
Op welke manieren de computer kan worden ingezet om de stofjes te testen staat beschreven in **Hoofdstuk 2**. Kortom zijn er twee hoofdmethodes: 1) het gebruik van 'machine learning' waarbij informatie wordt gebruikt van reeds geteste stofjes en 2) het gebruik van driedimensionale (3D) eiwitmodellen. Beide methodes worden in het algemeen gebruikt om de bioactiviteit van een stofje voor een eiwit te voorspellen. De bioactiviteit van een stofje weergeeft hoe goed het stofje zal binden aan een eiwit. Een stofje kan dus een andere bioactiviteit hebben voor eiwit A dan voor eiwit B. Heeft een stofje een hele lage bioactiviteit, dan word het stofje als inactief beschouwd voor het betreffende eiwit. Met de eerste methode, machine learning, wordt de computer getraind om patronen te herkennen in data. Vervolgens kan het gegenereerde model worden toegepast op nieuwe data om deze een waarde toe te kennen aan de hand van eerder geleerde trends. Dit klinkt abstract, maar in principe bestudeert het computermodel in zeer korte tijd duizenden stofje-eiwit combinaties. Hierdoor weet het welke stofjeseigenschappen (afmetingen, oplosbaarheid, etc.) het beste geschikt zijn voor een bepaald eiwit. Als er vervolgens een nieuw stofje door het model wordt gehaald, kan het model aan de hand van wat het al heeft gezien dus zeggen wat de bioactiviteit van het nieuwe stofje zal zijn. Om een machine learning model te kunnen trainen moet er dus op voorhand al informatie beschikbaar zijn van reeds in het lab geteste stofje-eiwit combinaties. Dit zijn namelijk "de feiten" waar het model conclusies op baseert. Een nieuw stofje dat door het model een voorspellende waarde wordt toegewezen hoeft uiteraard niet alvorens op bioactiviteit te zijn getest in het lab. Het voordeel van een machine learning model is dat het relatief snel is in vergelijking met 3D eiwitmodellen. Hoewel je dus wel voldoende data nodig hebt voordat je een nuttig model kan maken, kan het model als het eenmaal getraind is toe worden gepast op oneindig veel stofjes. De tweede modeleringstechniek om potentiële medicijnen mee te kunnen vinden is het simuleren van stofje-eiwit combinaties in 3D. Hierbij wordt er een 3D eiwitvisualisatie gemaakt en worden de stofjes er één voor één ingestopt. Hoe beter het stofje in het eiwit past, hoe hoger de bioactiviteit. Hoewel deze methode in het algemeen langzamer is dan machine learning modellen, is het vaak wel nauwkeuriger. Dit komt doordat de informatie van het eiwit in deze modellen meegenomen wordt om de bioactiviteit te voorspellen.

In **Hoofdstuk 3** wordt de machine learning methode toegepast om medicijnen te vinden voor de behandeling van diabetes type II. Een populair aangrijpingspunt voor de behandeling van deze ziekte is een transporteiwit genaamd SGLT2. SGLT2 bevindt zich in de nieren en is verantwoordelijk voor het terug transporteren van suikers in de urine naar het bloed. In een gezonde situatie is dit een handig systeem van ons lichaam dat ervoor zorgt dat er geen nuttige voedingstoffen verloren gaan. Echter, in het geval van diabetes type II zijn de bloedsuikerwaarden te hoog. Door SGLT2 met een medicijn te remmen, wordt er minder suiker terug het bloed in gestopt wat gunstige gevolgen heeft voor de bloedsuikerspiegels. Helaas hebben deze SGLT2 remmers op de langdurige termijn wel nadelige effecten als gevolg van de verhoogde suikerniveaus in de nieren. Bovendien moet het medicijn dat als een pilletje kan worden ingenomen, het hele lichaam doorreizen om bij zijn bestemming te komen. Daarom richt **Hoofdstuk 3** zich op het vinden van stofjes die binden op een alternatief doeleiwit om de bloedsuikerspiegels te verlagen. Het doeleiwit in dit hoofdstuk lijkt heel erg op SGLT2 en heet dan ook SGLT1. Hoewel SGLT1 zich ook in de nieren bevindt, is het ook aanwezig in de dunne darm. Omdat SGLT1 en SGLT2 heel erg op elkaar lijken en ook een soortgelijke functie hebben, zijn ze in staat om ook ongeveer dezelfde soort stofjes te binden. Daardoor is het lastig om een stofje selectief te maken voor SGLT1. In deze situatie is het echter niet zozeer nodig om een stofje selectief te maken voor

SGLT1 ten opzichte van SGLT2. Doordat SGLT1 zich ook in de dunne darm bevindt kan je ervoor kiezen om een stofje selectief maken voor de darmen waardoor het niet in staat is om SGLT1 en SGLT2 in de nieren te bereiken. Op deze manier zou je dus de suikeropname vanuit de darmen naar het bloed kunnen beperken en tegelijkertijd nierschade kunnen voorkomen. Een stofje selectief maken voor de darmen is mogelijk doordat je er eigenlijk voor zorgt dat het stofje het lichaam niet inkomt. Het maagdarmkanaal behoort namelijk nog tot ons uitwendig milieu. Je zou de darmen kunnen vergelijken met een groot filter: alles wat er doorheen komt is opgenomen door het lichaam en bevindt zich in het inwendig milieu. In het geval van een darm-selectief stofje zou dit kunnen betekenen dat het stofje te groot is om door het filter te passen. In **Hoofdstuk 3** wordt er met machine learning gezocht naar allerlei verschillende soorten stofjes die SGLT1 kunnen remmen. Deze stofjes dienen als een mooi startpunt voor het maken van selectieve remmers. Er wordt in het hoofdstuk vooral aandacht besteed aan het vinden van diverse stofjes – stofjes die niet veel op elkaar, of op bekende SGLT remmers lijken. Dankzij deze diversiteit is de mogelijkheid om een stofje te vinden dat darm-selectief gemaakt kan worden namelijk groter.

In **Hoofdstuk 3** ligt de focus op selectiviteit door een medicijn alleen te laten gaan waar het werkzaam moet zijn. Echter, in de meeste gevallen is een dergelijke afbakening van het verspreidingsgebied van een medicijn niet mogelijk. Hetzelfde geldt voor de casus die wordt besproken in **Hoofdstuk 4**. Dit hoofdstuk omvat twee eiwitten die bij veel processen in ons lichaam betrokken zijn: de A_1 en A_{2A} adenosine receptoren. Hoewel deze eiwitten gelijkenissen vertonen, hebben ze een tegenovergestelde werking. De één verlaagt het niveau cAMP in de cel en de ander verhoogt juist de hoeveelheid cAMP. De rol van cAMP is het doorgeven van signalen in de cel. Hierdoor worden andere eiwitten aangestuurd en ontstaat er een signaleringscascade. Met machine learning wordt er in **Hoofdstuk 4** een model gemaakt dat stofjes kan onderscheiden voor de A_1 en A_{2A} adenosine receptoren. In plaats van zoals op de traditionele manier de bioactiviteit van stofjes op beide eiwitten te voorspellen, voorspelt dit model de selectiviteit tussen de twee eiwitten. Vervolgens kunnen de stofjes gekozen worden die aan het gewenste profiel voldoen: A_1-selectief, A_{2A}-selectief of actief op beide eiwitten. **Figuur 1** laat zien hoe het verschil in bioactiviteit zich vertaalt naar selectiviteit. Hoewel het model in staat is verschillen in selectiviteit te detecteren, houdt het echter geen rekening met algehele bioactiviteit. Zo zou het kunnen dat een stofje A_1-selectief wordt voorspeld, maar eigenlijk helemaal niet bioactief is op de A_1 adenosine receptor (zoals het onderste voorbeeld in **Figuur 1**). Daarom is het selectiviteitsmodel gecombineerd met een 3D eiwitstructuur aanpak. Met 3D eiwitmodellen van zowel de A_1 en A_{2A} adenosine receptor konden de stofjes worden gepast in beide eiwitten. De stofjes die A_1-selectief werden voorspeld, maar niet in het 3D model van de A_1 adenosine receptor pasten konden er nu uitgefilterd worden. Andersom geldt dit natuurlijk ook voor stofjes die selectief zijn voor de A_{2A} adenosine receptor.

In **Hoofdstuk 5** wordt er ook gebruik gemaakt van zowel machine learning en 3D eiwitmodellen. Deze worden echter op een andere manier toegepast dan in **Hoofdstuk 4**, namelijk om het bioactiviteitsprofiel voor meerdere eiwitten te voorspellen in plaats van de selectiviteit tussen twee eiwitten. De eiwitten in **Hoofdstuk 5**, de zogenaamde kinases, zijn veelal betrokken bij complexe ziektes zoals kanker. De kaft van dit proefschrift weergeeft een vis die tegen de school inzwemt. Dit illustreert een eiwit dat zich anders gedraagt dan zou moeten, waardoor er een ziektebeeld ontstaat. In het geval van kanker zwemmen er meerdere vissen de verkeerde



Figuur 1. Verschillen in bioactiviteit weerspiegelen de selectiviteit tussen de A_1 en A_{2A} adenosine receptoren. Hoe groter de blauwe balk, hoe hoger de bioactiviteit. De minimaal benodigde hoeveelheid bioactiviteit om een effect te bereiken is weergeven in geel.

kant uit. Er zijn dan meerdere medicijnen nodig om elke vis te corrigeren. Het nadeel van het gebruik van meerdere medicijnen tegelijkertijd is dat deze medicijnen elkaar in de weg kunnen zitten, waardoor er ongewenste effecten ontstaan. **Hoofdstuk 5** beschrijft de zoektocht naar één medicijn dat meerdere eiwitten, of vissen, tegelijkertijd kan aansturen zonder de 'normale' vissen te beïnvloeden. Hiervoor worden meerdere machine learning en 3D eiwitstructuur modellen gebruikt om de bioactiviteit van stofjes te voorspellen voor een hele set kinases. Als resultaat krijg je een voorspelt bioactiviteitsprofiel voor elk stofje dat je door de modellen haalt. Vervolgens kunnen de stofjes met het meest gunstige profiel worden geselecteerd voor verder (lab)onderzoek.

Hoofdstuk 6 gaat in op een concept dat veel invloed kan hebben op het modelleren van selectiviteit, namelijk de bindingplaats van een stofje op een eiwit. Een eiwit heeft een hoofdbindingplaats waar een lichaamseigen stofje aan kan binden. Deze hoofdbindingplaats is vaak ook de plek waar een medicijn aan bindt. Echter, een eiwit kan ook meerdere bindingplaatsen hebben, welke tevens geschikt zouden kunnen zijn voor een medicijn om interactie mee aan te gaan. Vaak hebben de bindingplaatsen op een eiwit andere eigenschappen (bijvoorbeeld qua grote) wat betekent dat er verschillende soorten stofjes binden aan de verschillende bindingplaatsen. Als al deze stofjes in een computermodel worden gestopt zonder onderscheid te maken tussen de bindingplaatsen, kan dit de resultaten negatief beïnvloeden. **Hoofdstuk 6** laat zien dat wanneer de bindingplaatsinformatie meegenomen wordt in een machine learning model, dit ten gunste is van de nauwkeurigheid van voorspellingen. Helaas is bindingplaatsinformatie nog niet makkelijk

te verkrijgen en kan de verbetering dus niet zomaar worden doorgevoerd. Echter, met het 'text-mining' protocol dat in **Hoofdstuk 6** wordt beschreven kunnen de eerste stappen worden gezet naar het classificeren van de bindingplaatsen van stofjes. Door middel van text-mining kan de bindingplaatsinformatie namelijk uit wetenschappelijke publicaties worden gehaald, waarna het gebruikt kan worden om de voorspellingen van modellen te verbeteren.

Zoals in deze samenvatting is geschetst, zijn er meerdere manieren om de selectiviteit van medicijnen te modelleren en te voorspellen. Elke casus vraagt om een andere en eigen aanpak. Hoewel er dus geen eenduidige lijn is te trekken over hoe selectiviteit het beste gemodelleerd kan worden, valt er wel wat te zeggen over wat het succes van computermodellen bepaald. **Hoofdstuk 7** beschrijft onder meer dat de hoeveelheid data (informatie over bekende stofjes) belangrijk is, met name voor machine learning modellen. Deze modellen moeten namelijk een patroon leren herkennen en als er niet genoeg data is om van te leren worden de voorspellingen minder nauwkeurig. Bovendien speelt de diversiteit van data ook een rol. Een grote variabiliteit aan stofjes geeft meer kennis over wat wel en wat niet past in een eiwit. Het is dus heel nuttig om niet alleen veel data te hebben, maar ook om data beschikbaar te hebben waar veel informatie uitgehaald kan worden.

De potentiële medicijnen die met computermodellen worden geselecteerd worden eerst verder onderzocht voordat ze als medicijn beschikbaar komen. Hoewel we met computermodellen tijd kunnen winnen, duurt dit proces gemiddeld 10-15 jaar. Dit komt doordat we er zeker van moeten zijn dat de medicijnen echt veilig zijn voordat ze op de markt komen. Na het simuleren van stofjes op de computer, worden de stofjes getest in het lab. Vervolgens worden de meest succesvolle stofjes getest op dieren en uiteindelijk op vrijwilligers. Het kan zo zijn dat er ondanks het zorgvuldige werk er ergens in dit proces toch nog ernstige onvoorziene bijwerkingen aan het licht komen. Op het moment dat deze worden ontdekt betekent dit vaak het einde van de ontwikkeling van het specifieke stofje. Dat is natuurlijk zonde. Daarom is het blijven ontwikkelen van modellen die accurate voorspellingen kunnen doen van groot belang.

Glossary

Glossary

Affinity	Measure for the strength of substances to combine with each other (e.g. ligand-protein combination).
Allosteric	A non-orthosteric binding pocket on a protein.
Amino acids	The 'building blocks' of a protein or peptide.
Binding pocket	An area on a protein where substances (e.g. ligands, substrates) can bind.
Binding pose metadynamics	An automated, enhanced sampling method that assesses the compound stability in the binding pocket. A higher mobility indicates a less stable compound pose.
Bioactivity	A measure of effect caused by a compound when bound to a protein.
Clustering	Defining subgroups in a larger group (of molecules) based on structural similarity.
Compound	A (chemical) substance or small molecule.
Cryo-electron microscopy (cryo-EM)	A technique to determine the arrangement of atoms using cooled/frozen samples and electron microscopy.
Crystallography	A technique to determine the arrangement of atoms using crystals of the sample.
Protein crystallography	The determination of the arrangement and bonding of atoms of a protein using crystallography.
X-ray crystallography	A technique to determine the arrangement of atoms using crystals of the sample and X-rays refraction.
Descriptors	Identifiers used to describe a compound/protein; often chemical properties.
Docking	A computational technique to virtually place compounds into a rigid binding pocket of a protein.
Induced-fit docking	Similar to docking, but with (restricted) movement of the binding pocket amino acids enabled.
Ensemble scoring	A scoring method that uses multiple values and/or methods to score, e.g., the binding affinity of a compound.
Enzyme	A protein that carries out a biochemical reaction.

Homology model	A (protein) model that is based on a similar or related structure.
Inhibition	The opposite of activation; the action of inhibiting a process.
Ligand Endogenous ligand	A compound that interacts with or binds to a protein. The natural ligand that binds to the protein.
Machine learning	A computational technique that applies algorithms and statistical modeling to automatically learn (patterns) from data.
Molecular dynamics simulation	A (newtonian) physics-based technique to simulate the movements of atoms and molecules.
Orthosteric	The binding pocket on a protein where the endogenous ligand or substrate binds.
Pathophysiology	The disordered physiological processes associated with disease or injury.
Polypharmacology	Compounds that act on multiple proteins.
Protein	Large molecules composed of chains of amino acids of which the structure is encoded in the DNA.
Protein expression	The biologic synthesis of proteins.
Proteochemometrics (PCM)	Supervised machine learning of bioactivity based on explicit molecular and target descriptors (typically proteins).
Quantitative structure-activity relationship (QSAR)	The relation between the chemical structure of a compound and its activity on a protein.
Random forest	An ensemble learning method that applies multiple unpruned decision trees.
Receptor	A membrane bound protein that can activate second messengers after ligand recognition.
Selectivity	The property of affecting (e.g. inhibiting) a specific protein, while not affecting others at a similar level.
Statistical modeling	A modeling technique based on mathematics. The model is defined by an equation that approximates the data.

Structural protein-ligand interaction fingerprints (SPLIFs)[i]	An interaction fingerprint that explicitely encodes the interaction between ligand and protein in a structural way and therefore implicitly captures all types of ligand–protein interactions.
Substrate	A molecule that is used by enzymes in catalysis.
Target	A protein considered in ligand binding.
Anti-target	A protein to which ligand binding should be avoided.
Off-target	A protein that will result in undesired effects upon ligand binding.
On-target	A protein that will result in desired effects upon ligand binding.
T-Distributed Stochastic Neighbor Embedding (t-SNE)	A non-linear technique for dimensionality reduction to visualize high-dimensional data.
Text mining	Deriving data from text by automatically searching the text using the computer.
Transporter	A protein that transports substances across the cell membrane.
Virtual screening	A computational way of testing compounds against a protein.

i Da, C.; Kireev, D. Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. J. Chem. Inf. Model. 2014, 54, 2555-2561.

List of Publications

List of Publications

Burggraaff, L.; Oranje, P.; Gouka, R.; van der Pijl, P.; Geldof, M.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. *Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 Using Proteochemometric Modeling*. Journal of Cheminformatics **2019**, 11, 15. https://doi.org/10.1186/s13321-019-0337-8.

Burggraaff, L.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. *Quantitative Prediction of Selectivity between the A_1 and A_{2A} Adenosine Receptors*. Journal of Cheminformatics **2020**, 12, 33. https://doi.org/10.1186/s13321-020-00438-3

Burggraaff, L.; Lenselink, E. B.; Jespers, W.; van Engelen, J.; Bongers, B. J.; Gorostiola González, M.; Liu, R.; Hoos, H. H.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. *Successive Statistical and Structure-Based Modeling to Identify Chemically Novel Kinase Inhibitors*. Journal of Chemical Information and Modeling **2020**. https://doi.org/10.1021/acs.jcim.9b01204.

Burggraaff, L.; van Veen, A.; Lam, C. C.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. *Annotation of allosteric compounds to enhance bioactivity modeling for class A GPCRs*. submitted.

Sydow, D.*; Burggraaff, L.*; Szengel, A.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P.; Volkamer, A. *Advances and Challenges in Computational Target Prediction*. Journal of Chemical Information and Modeling **2019**. https://doi.org/10.1021/acs.jcim.8b00832.
*these authors contributed equally

Oranje, P.; Gouka, R.; Burggraaff, L.; Vermeer, M.; Chalet, C.; Duchateau, G.; van der Pijl, P.; Geldof, M.; de Roo, N.; Clauwaert, F.; Vanpaeschen, T.; Nicolaï, J.; de Bruyn, T.; Annaert, P.; IJzerman, A. P.; van Westen, G. J. P. *Novel Natural and Synthetic Inhibitors of Solute Carriers SGLT1 and SGLT2*. Pharmacology Research & Perspectives **2019**, 7, e00504. https://doi.org/10.1002/prp2.504.

Hameed, D. S.; Sapmaz, A.; Burggraaff, L.; Amore, A.; Slingerland, C. J.; van Westen, G. J. P.; Ovaa, H. *Development of Ubiquitin-Based Probe for Metalloprotease Deubiquitinases*. Angewandte Chemie International Edition **2019**, 58, 14477–14482. https://doi.org/10.1002/anie.201906790.

Ruano-Ordás, D.; Burggraaff, L.; Liu, R.; van der Horst, C.; Heitman, L. H.; Emmerich, M. T. M.; Mendez, J. R.; Yevseyeva, I.; van Westen, G. J. P. *A Multiple Classifier System Identifies Novel Cannabinoid CB2 Receptor Ligands*. Journal of Cheminformatics **2019**, 11, 66. https://doi.org/10.1186/s13321-019-0389-9.

Koenders, S. T. A.; Wijaya, L. S.; Erkelens, M. N.; Bakker, A. T.; van der Noord, V. E.; van Rooden, E. J.; Burggraaff, L.; Putter, P. C.; Botter, E.; Wals, K.; van den Elst, H.; den Dulk, H.; Florea, B. I.; van de Water, B.; van Westen, G. J. P.; Mebius, R. E.; Overkleeft, H. S.; Le Dévédec, S. E.; van der Stelt, M. *Development of a Retinal-Based Probe for the Profiling of Retinaldehyde Dehydrogenases in Cancer Cells*. ACS Central Science **2019**. https://doi.org/10.1021/acscentsci.9b01022.

Yang, X.; Dilweg, M.; Osemwengie, D.; Burggraaff, L.; Es, D.; Heitman, L.; IJzerman, A. *Design, Synthesis and Pharmacological Profile of LUF7746, a Novel Covalent Partial Agonist for the Adenosine A_1 Receptor*. **2020**. https://doi.org/10.26434/chemrxiv.11903286.

Zhou, J.; Mock, E. D.; Martella, A.; Kantae, V.; Di, X.; Burggraaff, L.; Baggelaar, M. P.; Al-Ayed, K.; Bakker, A.; Florea, B. I.; Grimm, S. H.; den Dulk, H.; Li, C. T.; Mulder, L.; Overkleeft, H. S.; Hankemeier, T.; van Westen, G. J. P.; van der Stelt, M. *Activity-Based Protein Profiling Identifies α-Ketoamides as Inhibitors for Phospholipase A2 Group XVI*. ACS Chemical Biology **2019**. https://doi.org/10.1021/acschembio.8b00969.

Zhou, J.; Mock, E. D.; Al Ayed, K.; Di, X.; Kantae, V.; Burggraaff, L.; Stevens, F.; Martella, A.; Mohr, F.; Jiang, M.; van der Wel, T.; Wendel, T. J.; Ofman, T.; Tran, Y.; de Koster, N.; van Westen, G. J. P.; Hankemeier, T.; van der Stelt, M. *Structure-Activity Relationship Studies of α-Ketoamides as Inhibitors of the Phospholipase A and Acyltransferase (PLAAT) Enzyme Family*. Journal of Medicinal Chemistry **2020**. https://doi.org/10.1021/acs.jmedchem.0c00522.

Mladic, M.; de Waal, T.; Burggraaff, L.; Slagboom, J.; Somsen, G. W.; Niessen, W. M. A.; Manjunatha Kini, R.; Kool, J. *Rapid Screening and Identification of ACE Inhibitors in Snake Venoms Using At-Line Nanofractionation LC-MS*. Analytical and Bioanalytical Chemistry **2017**, 409, 5987–5997. https://doi.org/10.1007/s00216-017-0531-3.

Curriculum Vitae

Curriculum Vitae

Lindsey Burggraaff was born on the 17th of November 1990 in Purmerend, The Netherlands. After graduating from pre-university education at the Da Vinci College in Purmerend in 2010, she started her study in pharmaceutical sciences at the VU in Amsterdam. After obtaining her bachelor degree, she continued with her master in Drug Discovery and Safety. She completed the curriculum of two specializations: "Drug Discovery and Target Finding" and "Computational Medicinal Chemistry and Toxicology". Her master studies were finalized with an internship in the computational medicinal chemistry group where she modeled selectivity of compounds between human and viral phosphodiesterases in the framework of medicines against neglected parasitic diseases.

In 2015 she pursued her interest in computational modeling by performing a PhD at Leiden University, during which she became an expert in data science and 3D modeling of ligand-protein interactions. Using machine learning, she successfully modeled glucose transporter proteins for a project with Unilever. She provided computational expertise to different fields of research and closely collaborated with multiple institutes including: the department of Cell and Chemical Biology, Oncode Institute, Leiden University Medical Center; the Institute of Physiology, Charité Universitatsmedizin, Berlin, Germany; and Karlsruhe Institute of Technology, Karlsruhe, Germany. Moreover, she visited the Cell and Molecular Biology group in Uppsala, Sweden, as guest researcher to master additional 3D modeling techniques.

Lindsey frequently presented her work at national and international conferences. She presented at the Centre for Computational Life Sciences, Leiden; EuroQSAR, Verona, Italy; ULLA, Helsinki, Finland; and the FIGON Dutch Medicines Days, Ede, of which the latter was awarded with the best poster award. Moreover, she obtained a travel grant to support her attendance at the Eight Joint Sheffield Conference on Chemoinformatics in Sheffield, UK, where she presented her research to experts in the field. Next to scientific outreach, she focused on conveying her research to a broader audience - the general public. With her participation in FameLab, a renowned international science communication competition, she reached the national final with a pitch that was praised for its clarity. Furthermore, she presented her research to a laymen audience at "the night of discoveries", a yearly cultural event in Leiden that showcases the wonders of science to thousands of visitors.

She continues her career in digital acceleration and currently takes on new data-driven challenges in industry.

Acknowledgements

Acknowledgements

This thesis is a result of four years of research, during which I met many inspiring people and gratefully experienced the compassion of many colleagues and peers. I would like to thank everyone who contributed to this thesis in any way, be it scientifically or socially.

Thank you, Ad and Herman for your guidance, advice and kindness during my research – I am grateful to have had you as my promotors. Ad, I appreciate our scientific and social interactions. Herman, thank you for your critical eye and your refreshing opinions from an industry point of view.

Gerard, my co-promotor, thank you for always being there to provide me with feedback on my scientific work. I am thankful for the enjoyable social activities we shared. You made me feel trusted and appreciated.

Thank you to all my collaborators from the other faculties, industry and institutes, for providing me with the opportunity to broaden my perspective on the applicability of computational methods in other research fields. Michael, thank you for you advice on computer science matters.

Furthermore, I would like to thank all my (former) colleagues who contributed to the pleasant working environment. Especially my computational colleagues Bart, Willem, Shardul and Olivier. Bart, the last time I checked there was still a physical trace at our office wall reminding me of your 4 o'clock grapefruit time. May it live on through many generations of PhDs. Willem, you are awesome and you know it! I want to point out to Shardul that the key you broke is still blamed on me. Being a boxer has not been in my favour, especially when the window spontaneously cracked shortly after (Ad, I swear it was just coincidental). Nevertheless, you are my friend in business and beyond. Olivier, you are truly one of a kind and the most soulful dancer I have ever met; you are glorious.

Thank you lab-ladies Anna, Xue, Natalia, Rongfang, Xuesong and Tasia, who made coffee breaks interestingly fun and with whom I experienced many social events outside of working hours. We had a lot of fun together and I will definitely miss working with you.

To the students who performed their research internships under my supervision: thank you for your hard work and for making me the supervisor I am today. This goes especially for my master students Amber, Jenke, Florian, Marina and Martijn. Although your personalities could not differ more, I enjoyed supervising every single one of you. You are great minds and I surely believe you will do great in your future careers.

My paranymphs, Nadine and Willem, thank you for your support and for being by my side on such a special day. Nadine, thank you for being such a dear friend over the past years, and for the many more years to come! Willem, you were a valued office neighbour and are an even more valued friend. With you, laughter can always be found right around the corner.

Special thanks to my parents, Hans and Ellen, of whom I received unconditional love and support. Sandy, Dana, Narita, Bradley and Rozemarijn, I am grateful for the bond we have - I can praise myself lucky with such amazing siblings.