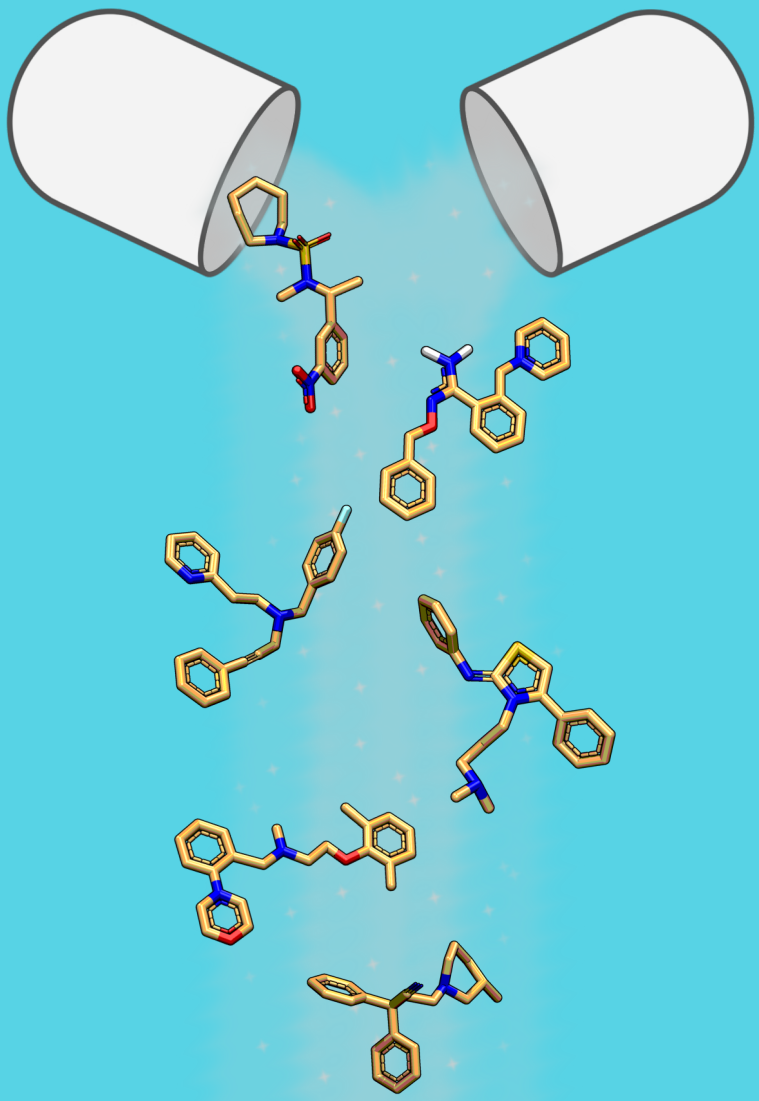
Proteins in harmony: Tuning selectivity in early drug discovery

Burggraaff, L.

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 Type1 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.