Transforming data into knowledge for intelligent decision-making in early drug discovery

Paricharak, S.A.

Chapter seven

General Conclusions

Conclusions from this thesis

Pharmaceutical research and development is plagued notoriously by high cost and substantial drug attrition rates. It is a field surrounded by much uncertainty, caused by poor understanding of the sheer complexity of disease states and drug efficacy at both the pre-clinical and clinical stages.[1] Not only have academia and the industry performed research independently in an attempt to ameliorate this, but they have also joined forces in the form of research collaborations and public releases of biological and chemical data from the industry.[2] The latter allowed academics to contribute to solving problems acknowledged as challenging by the industry. During my PhD, I embarked on a journey to analyze publicly available industrial data, and was fortunate to continue on to more hands-on collaborations with Novartis, gaining access to large-scale proprietary data and experiencing current trends in big pharmaceutical companies from the inside out.

At the top-most level, the goal of my PhD was to improve drug attrition rates and reduce research costs. At a lower level, my objective was to use computational methods to improve the efficiency of early-stage drug discovery efforts where typically millions of compounds are tested, to subsequently only select a very small subset for further investigation.[3] These methods entail the use of existing bioactivity information to build computational models to anticipate compound activity *in silico* (bioactivity modeling),[4] thereby providing more promising starting points for testing compared to random selection. Ample room for improvement is envisaged given the relatively low hit rates in many early-stage drug discovery campaigns,[5] and I hope my research will lead to savings in time and resources.

In this thesis, I inspected bioactivity modeling from various angles. The performance of a computational model directly depends on the quality and relevance of the experimental data it is trained on. Ideally, predictive power over a diverse set of compounds (i.e., a set containing many structurally dissimilar compounds) is desired, and therefore, selecting diverse sets for model learning is crucial for overall performance. However, chemical space is vast, as over 10^{63} small molecules with a mass comparable to many drugs possibly exist,[6] making selection and testing of a large fraction for model building unfeasible. This highlights the need for efficient exploration of chemical space for model building.

In *Chapter three*, I described the relevance of chemical space for drug discovery and discuss existing methods for its effective sampling. Chemical diversity is an ambiguous concept that depends on the set of characteristics (descriptors) used to compare molecules. Examples of such descriptors include molecular shape, atom connectivity, solubility and charge amongst many others. In this study, the examination of a wide range of commonly used descriptors across chemical libraries varying in size and diversity led to insights into correlations between descriptors in terms of (1) diversity assessment and (2) retrieval of active compounds from ChEMBL,[7] a public bioactivity database. In other words, the results from this chapter provide a perspective on the ambiguity of the concept of molecular diversity, and come with practical examples of the use of common descriptors for the selection of activity-enriched starting points. It is hoped that the reader realizes how strongly the results obtained (i.e., chemical space sampled) depend on the descriptor used for diversity analysis.

While *Chapter three* represents a reflective analysis on state-of-the-art diversity assessments of chemical libraries, *Chapter four* illustrates the firsthand application of a computational method geared toward systematic compound prioritization. The work described in *Chapter four* was performed at Novartis and was based on large-scale proprietary high-throughput screening (HTS) data. One of the key drawbacks of HTS campaigns performed routinely in the pharmaceutical industry is the high upfront cost in relation to the number of active compounds discovered.

This study addressed precisely this issue by proposing a new compound screening paradigm and comprehensively validating it for the first time on an unparalleled scale. The screening strategy involved the iterative selection of compounds chemically and biologically similar to actives identified in multiple rounds of screening, consistently leading to over tenfold increases in efficiency with respect to activity and diversity enrichments of selected compound sets. The results obtained from this strategy are illustrated in **Figure 34.**

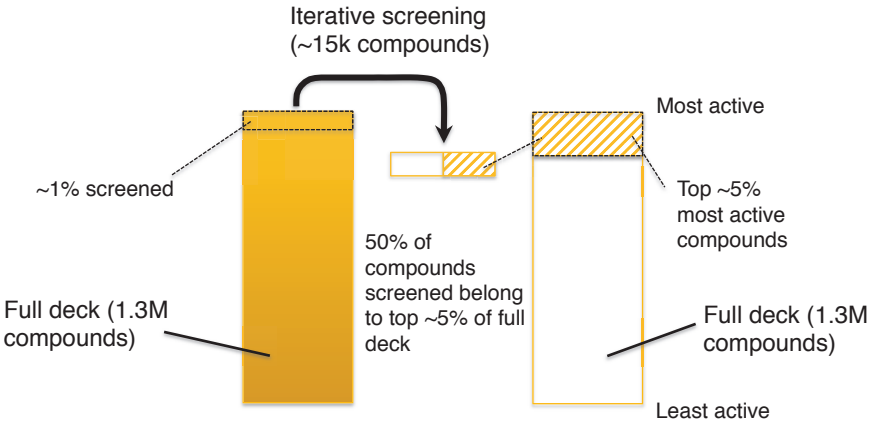

Figure 34. Illustration of efficiency gains using iterative screening. Half of the compounds selected iteratively, which corresponds to only 0.5% of the entire screening collection of 1.3 million compounds (full deck), were found to be among the top 5% of compounds in terms of activity. This indicates a tenfold enrichment in activity.

Overall, half of the compounds selected iteratively, which corresponds to only 0.5% of the entire screening collection of 1.3 million compounds (full deck), were found to be among the top 5% of compounds in terms of activity. A cornerstone for the success of this approach was the "high-throughput screening fingerprint" (HTS-FP),[8] which is a biological similarity metric that compares molecules not on basis of their structure, but of their effect on cells and targets, hereby harnessing the vast amount of biochemical information typically available at a pharmaceutical company for enhanced activity modeling.

Although promising and intuitive results were obtained in *Chapter four*, the approaches described therein were straightforward from a conceptual point of view. This led to the question: if simple approaches already resulted in promising results, how much more is there to gain from employing more sophisticated approaches? At the same time, infrastructure-related difficulties in HTS at Novartis created a demand for small pre-composed compound sets for facilitated routine exploratory screening. Inspired by the current wave of big data analytics and machine learning, I converged both aforementioned points and delved deeper into publicly available HTS data to derive an "informer compound set" using advanced machine learning approaches. Once screened, this set provided the most information on which compounds to test subsequently from the yet unscreened remainder of the collection,

regardless of biological target. *Chapter five* describes the results obtained for this study.

The final research chapter, *Chapter six*, examines bioactivity modeling in a different context: to predict the mode-of-action of a collection of anti-malarial compounds, published by GSK in an attempt to combat the lack of novel drugs for neglected diseases.[2] Despite the lack of target annotations, these compounds showed inhibitory effects on parasitic cell growth (phenotypic effects). In this study, the integration of two machine learning methods (Bayesian target prediction and proteochemometric modeling)[9] is illustrated, exploiting the advantages of both methods for simultaneous polypharmacology and affinity predictions.

Having investigated HTS data thoroughly both at Novartis and in the public domain, I realized not only the importance of HTS data in pharmaceutical research, but also the difficulty in the design of HTS campaigns and post-screen analysis. An undirected, random high-throughput search for active compounds can in some cases be compared to hunting for a needle in a haystack. Indeed, much effort is put in intelligent design of HTS at the compound library composition, post-screen analysis and bioactivity modeling stages. Of note, remarkable advancements have been made in bioactivity modeling fueled by the recent introduction of large-scale HTS-FP,[8] leading to enhanced hit rates and insights into compound mode-of-action.[3,10-12] In *Chapter two*, I reviewed data-driven approaches used in HTS, and elaborated on the recent rapid progress in bioactivity modeling, outlining its significance in the field.

Future perspectives

Drug discovery has witnessed numerous changes over the past decades. Below, I outline my perspective on cheminformatics analyses in HTS and bioactivity modeling.

With the exception of some academic screening centers,[13] HTS is primarily performed in the pharmaceutical industry due to the high cost and infrastructure requirements. As a consequence, although the PubChem[13] repository contains data for over two hundred HTS assays, the amount of public HTS data available is still limited compared to the amount generated in the pharmaceutical industry. This scarcity is also a reality for other types of (lower-throughput) bioactivity data. Additionally, consistency in quality is an issue for public (HTS) data due to the disparate sources it originates from.

This limits and complicates research in academia. During my PhD I repeatedly encountered the data paucity problem, caused either by the lack of (good quality) data or incompleteness in the data available. In light of this, I believe that academic endeavors should couple computational work more tightly with experimental validation, enabling efficient cycles of hypothesis generation, testing and feedback for improved overall output. In addition, active collaboration between academia and the pharmaceutical industry can to some extent mitigate the issues discussed.

However, I believe that at a higher level the primary purpose of academia is not only to document and distribute existing knowledge, but also to create new knowledge. Profit-driven corporations cannot necessarily afford this given the risk of no return on investment when no clear application is envisaged at outset. Therefore, academia should adjust its research scope accordingly and aim to generate an orthogonal stream of knowledge to that generated in the industry. Many partnerships between academia and the industry, including the National Center for Advancing Translational Sciences (NCATS),[14] the American Cancer society and knowledge exchange programs between big pharmaceutical companies and academia, among many others[15] enable translational and drug discovery research on a larger scale than ever before. Such partnerships have brought industrial expertise in areas such as assay development closer to the academic domain.[15]

Taking advantage of the opportunities offered in this setting, academic drug discovery could focus on high quality basic research aiming to understand the fundamentals of underexplored disease biology. Other opportunities for academic drug discovery include drug discovery for neglected diseases and drug repositioning. Often little incentive exists for these endeavors in the industry due to the limited market size (neglected diseases), and intellectual property issues around the original drug complicating commercialization (drug repositioning).[15] At a more fundamental level, poorly understood phenomena such as protein-protein binding could be examined in detail, supplemented by novel exploratory analyses of biochemical data aiming to investigate fresh high-risk concepts (e.g., bioactivity modeling based on deep learning, a machine learning method which has recently become popular) coupled with experimental validation.

Better understanding of the relationship between epigenetics and disease pathology has recently sparked interest in the field of epigenetic drug discovery: a number of partnerships have formed in the quest for epigenetic

drugs, such as the one between GlaxoSmithKline and Cellzome for combating immune and inflammatory disorders.[16] Another significant effort is the public-private partnership led by the Structural Genomics Center and involving GlaxoSmithKline and other institutes. The goal of this effort was to generate well-defined 'chemical probes' for epigenetic targets based on potency, selectivity, and cellular engagement requirements, and to release these probes into the public domain.[17] This field represents an exciting opportunity for academic drug discovery, as even though recent progress in epigenetics has given hope for novel drug discovery, the field is still immature, and it is likely that thorough research will lead to novel insights.[18]

The industry should in turn focus on application-driven research, as existing resources and infrastructure facilitate this. Here, the aim should be to use public and proprietary knowledge to accelerate and improve drug discovery (e.g., improving HTS efficiency, and integrating diverse data, such as transcriptomics, metabolomics and genomics data for new insights) with the ultimate aim of inventing effective therapies. An interesting study by Wassermann *et al.*[19] illustrated the concept of dark chemical matter, where sets of compounds that were consistently identified as inactive across a wide range of assay biology occasionally contained potent hits with selective activity and clean safety profiles. These compounds were found to be valuable starting points for further research. Exploring this finding thoroughly could provide further avenues into understanding the characteristics of bioactive molecules, which is an opportunity for the pharmaceutical industry.

While collaborations between academia and industry are certainly useful and have their own place, my intention is to alert the reader to the original scope of academia: to generate knowledge and tools (e.g., software) that are not severely subject to profit-driven interests. In conjunction with the result-oriented approach of the industry, I believe that an overall good net result can be achieved in terms of research output and productivity, even if direct collaboration between academia and the industry is not necessarily enhanced. This concept, in some sense similar to *active learning* described in detail in *Chapter five*, is shown in **Figure 35**: academia explores uncharted territories of knowledge space, followed by application around these areas by the pharmaceutical industry.

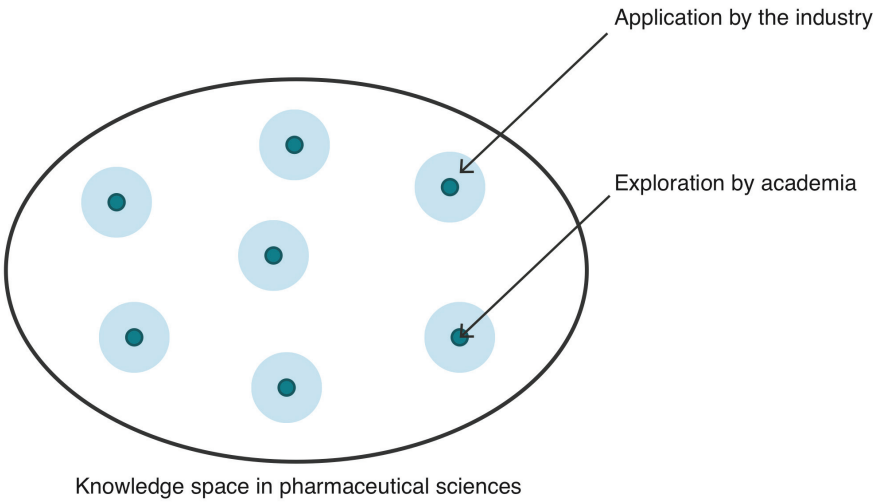

Figure 35. Exploration of knowledge space in pharmaceutical sciences by collaborative efforts of academia and the industry. Academia could focus on research areas that are not necessarily profit-driven (e.g., neglected diseases and drug repositioning) and therefore more likely to be neglected by the pharmaceutical industry. The industry should in turn focus on application-driven research with the ultimate aim of inventing effective therapies.

My industrial placement has provided me with exposure to recent trends in cheminformatics. Of note, I observed a resurgence in phenotypic screening, where collections of compounds are tested for desirable effects on cells and/or tissues without upfront knowledge on their mode-of-action.[20] Upon discovery of active compounds, effort is put into elucidating their mode-of-action, prompting new approaches for target identification. HTS-FP[8] was developed in this regard and proved to be a foundation for a body of work[3,5,10,11,21] on bioactivity modeling leading to increased efficiency and novel insights into compound mode-of-action. This work was published in quick succession due to the remarkable improvements HTS-FP offered over conventional structural similarity metrics. The key novelty about HTS-FP is the dimensions in which it compares molecules. While chemical fingerprints relate compounds on the basis of their structure, HTS-FP describes compounds based on their bioactivity across a large number of biologically relevant end points, including activity against target proteins and phenotypic effects. Hence, the partially incorrect implicit assumption that chemical similarity correlates with similar activity patterns is circumvented, and at the same time empirical data is used to define similarity in a directly relevant dimension.

Drawing from the aforementioned trends in the use of phenotypic screening and bioactivity-based similarity metrics, supported by many promising results from recent studies,[3,5,10,11,21] I believe that many exciting discoveries remain to be made by examining compounds from a biologically relevant point of view. While much of the low-hanging fruit (efficiency gains, mode-of-action analyses) has already been picked, an in-depth analysis of activity correlations across independent biological end points (i.e., cells, tissues and protein targets) has not been performed. It is my firm conviction that this analysis represents an opportunity potentially leading to unmapped insights into bioactivity-based similarities between proteins. If these analyses prove useful, novel insights into phylogenetically non-related proteins similar in bioactivity space could further improve modeling efforts, for example by enabling proteochemometric modeling[9] across a more diverse range of proteins.

Final remarks

This thesis describes various aspects of bioactivity modeling in drug discovery, and touches upon the relevant topics of efficient exploration of chemical space, and different ways of improving screening efficiency in HTS campaigns. Given sufficient high quality experimental data, bioactivity modeling is relatively inexpensive and at the same time has the potential of providing promising starting points for experimental validation. This is of great value to drug discovery, a field with much uncertainty and substantial drug attrition rates. Looking back at the past four years, I feel fortunate to have had an opportunity to make contributions to bioactivity modeling for more intelligent decision-making in early (academic) drug discovery, and sincerely hope that my work leads to novel ideas in the future…

References

(1) Carter, G. T. (2011) Natural products and Pharma 2011: Strategic changes spur new opportunities. *Nat. Prod. Rep. 28*, 1783–1789.

(2) Gamo, F.-J., Sanz, L. M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J.-L., Vanderwall, D. E., Green, D. V. S., Kumar, V., Hasan, S., Brown, J. R., Peishoff, C. E., Cardon, L. R., and Garcia-Bustos, J. F. (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature 465*, 305–310.

(3) Paricharak, S., IJzerman, A. P., Bender, A., and Nigsch, F. (2016) Analysis of iterative screening with stepwise compound selection based on Novartis in-house HTS data. *ACS Chem. Biol. 11*, 1255–1264.

(4) Koutsoukas, A., Lowe, R., KalantarMotamedi, Y., Mussa, H. Y., Klaffke, W., Mitchell, J. B., Glen, R. C., and Bender, A. (2013) In silico target predictions: defining a benchmarking

dataset and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window. J. Chem. Inf. Model.
(5) Maciejewski, M., Wassermann, A. M., Glick, M., and Lounkine, E. (2015) An Experimental Design Strategy: Weak Reinforcement Leads to Increased Hit Rates and Enhanced Chemical Diversity. J. Chem. Inf. Model. 55, 956–962.
(6) Koutsoukas, A., Lowe, R., KalantarMotamedi, Y., Mussa, H. Y., Klaffke, W., Mitchell, J. B., Glen, R. C., and Bender, A. (2013) In silico target predictions: defining a benchmarking dataset and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model. 53*, 1957-1966.
(7) Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl. Acids Res. 40*, D1100–1107.
(8) Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., Deng, Z., Davies, J. W., Jenkins, J. L., and Glick, M. (2012) Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol. 7*, 1399–1409.
(9) Van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., Van Vlijmen, H. W. T., and Bender, A. (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun. 2*, 16–30.
(10) Wassermann, A. M., Lounkine, E., and Glick, M. (2013) Bioturbo Similarity Searching: Combining Chemical and Biological Similarity To Discover Structurally Diverse Bioactive Molecules. *J. Chem. Inf. Model. 53*, 692–703.
(11) Wassermann, A. M., Lounkine, E., Urban, L., Whitebread, S., Chen, S., Hughes, K., Guo, H., Kutlina, E., Fekete, A., Klumpp, M., and Glick, M. (2014) A Screening Pattern Recognition Method Finds New and Divergent Targets for Drugs and Natural Products. *ACS Chem. Biol. 9*, 1622–1631.
(12) Riniker, S., Wang, Y., Jenkins, J. L., and Landrum, G. A. (2014) Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model. 54*, 1880–1891.
(13) Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., Bolton, E., Gindulyte, A., and Bryant, S. H. (2012) PubChem's BioAssay Database. *Nucl. Acids Res. 40*, D400–D412.
(14) Littman, B. H. (2011) An NIH National Center for Advancing Translational Sciences: is a focus on drug discovery the best option? *Nat. Rev. Drug Discov. 10*, 471.
(15) Dahlin, J. L., Inglese, J., and Walters, M. A. (2015) Mitigating risk in academic preclinical drug discovery. *Nat. Rev. Drug Discov. 14*, 279–294.
(16) DeWoskin, V. A., and Million, R. P. (2013) The epigenetics pipeline. *Nat. Rev. Drug Discov. 12*, 661–662.
(17) Brown, P. J., and Müller, S. (2015) Open access chemical probes for epigenetic targets. *Futur. Med. Chem. 7*, 1901–1917.
(18) Arguelles, A. O., Meruvu, S., Bowman, J. D., and Choudhary, M. (2016) Are epigenetic drugs for diabetes and obesity at our door step? *Drug Discov. Today 21*, 499–509.
(19) Wassermann, A. M., Lounkine, E., Hoepfner, D., Le Goff, G., King, F. J., Studer, C., Peltier, J. M., Grippo, M. L., Prindle, V., Tao, J., Schuffenhauer, A., Wallace, I. M., Chen, S., Krastel, P., Cobos-Correa, A., Parker, C. N., Davies, J. W., and Glick, M. (2015) Dark chemical matter as a promising starting point for drug lead discovery. *Nat. Chem. Biol. Chem. Biol. 11*, 958–966.
(20) Kotz, J. (2012) Phenotypic screening, take two. *SciBX 5*, 1–3.
(21) Paricharak, S., IJzerman, A. P., Jenkins, J. L., Bender, A., and Nigsch, F. Data-driven derivation of an "informer compound set" for improved selection of active compounds in high-throughput screening. *Submitted*