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

Paricharak, S.A.

Transforming Data into Knowledge for Intelligent Decision-making in Early Drug Discovery

Shardul Atul Paricharak

The research described in this thesis was performed at the following institutes: (1) Centre for Molecular Informatics at the Department of Chemistry, University of Cambridge (Cambridge, United Kingdom), (2) Novartis Institutes for BioMedical Research, Novartis Pharma AG (Basel, Switzerland), and (3) Division of Medicinal Chemistry of the Leiden Academic Centre for Drug Research, Leiden University (Leiden, The Netherlands). The research was financially supported by the Netherlands Organization for Scientific Research (NWO-017.009-065), Novartis Institutes for BioMedical Research, and the Prins Bernhard Cultuurfonds.

This thesis was printed by Haveka.

ISBN 978-3-033-06023-4

Transforming Data into Knowledge for Intelligent Decision-making 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 9 februari 2017
klokke 10 uur

door

Shardul Atul Paricharak
geboren te Willemstad, Curaçao in 1989

Promotor:	Prof. Dr. A.P. IJzerman
Co-promotor:	Dr. A. Bender (University of Cambridge, Verenigd Koninkrijk)
Promotiecommissie:	Prof. Dr. J.A. Bouwstra Dr. C. de Graaf (Vrije Universiteit Amsterdam, Nederland) Prof. Dr. J. Kirchmair (Universität Hamburg, Duitsland) Prof. Dr. M. Danhof Prof. Dr. J.N. Kok

Contents

Chapter one

General Introduction

About this thesis

This thesis describes various analyses of life science data resulting in new knowledge, which can be used to support decision-making in early drug discovery. The results obtained herein are envisaged to lead to efficiency gains in future experimental campaigns and novel insights into compound mode-of-action (i.e., the protein target modulated for the desired phenotypic effect). Below, I introduce the relevance of computational drug discovery, the increase in publicly available life science data creating opportunities for bioactivity modeling, and the role cheminformatics and bioinformatics play in the latter. In the last section of this chapter, I specifically outline the objectives of this thesis.

Background

As long as mankind exists, there is a need for medicines. Historically, compounds of natural origin (i.e., original natural products, products derived semi-synthetically from natural products, or synthetic products based on natural product models) were used to treat diseases.[1] Later, early drug discovery resulted from multidisciplinary research with key contributions from (medicinal) chemists, pharmacologists, and clinical scientists.[2,3] However, looking back today, drug discovery has come a long way. Advances in molecular biology have increased our understanding of many complex diseases, allowing for the design of drugs aimed at modulating key protein targets in addition to phenotypic testing alone.[3] At the same time, rapid improvements in combinatorial and parallel chemistry, and developments in assay technologies led to larger compound collections in the pharmaceutical industry and the testing thereof across a larger number of biological entities. Improvements in automation and robotics over the past decades even created the possibility of rapidly testing very large collections comprising 1–2 million compounds routinely (high-throughput screening – HTS), aiming to interrogate compound libraries in a brute-force manner to identify promising active molecules (hits).[4]

Taken together, the aforementioned progress led to the increased publication of diverse life science data, including bioactivity and phenotypic data, describing compound activity on protein targets and on cells or tissues, respectively. The idea of data sharing was further reinforced by the Wellcome Trust, which awarded £4.7 million to EMBL-EBI to support the acquisition and publication of data on drugs and drug-like molecules from Galapagos

NV.[5] The data from this acquisition formed the basis of ChEMBL,[6] a large-scale database which curates life science data for facilitated public access.

The availability of ever-growing amounts of unanalyzed data gives rise to a central question: how can we exploit this data, convert it to knowledge and support decision-making in drug discovery? Undoubtedly, many proposals ranging broadly from the analysis of novel data types, data integration, to the development of novel analysis methods answer this question. In this thesis, I primarily focus on bioactivity data modeling, aiming to anticipate compound activity *in silico*. Here, the intention is to save precious resources required for experimental testing by prioritizing compounds by their likelihood of being active, with the hope that such prioritization results in activity-enriched subsets of compounds.

A core assumption forms the basis of bioactivity modeling: similar molecules exhibit similar activity patterns (the principle of molecular similarity).[7] However, this assumption is partially incorrect, as evidenced by the existence of activity cliffs[8] that represent pairs of compounds that exhibit large differences in activity despite their perceived similarity. Additionally, similarity is an ambiguous concept and can be based on a number of compound signatures (descriptors), which characterize compounds in terms of their intrinsic properties. The choice of descriptor determines the relative positioning of compounds in descriptor space, which in turn directly defines their nearest and most distant neighbors. These descriptors can be based on chemical properties (e.g., molecular shape, atom connectivity, solubility and charge amongst many others),[9-12] or biological properties (i.e., activity profiles across a panel of relevant protein or cellular targets).[13-16] The key benefit of using the latter descriptor type is that the partially incorrect assumption that chemical similarity correlates with similar activity patterns is circumvented, while using empirical data to define similarity in the most relevant dimension (bioactivity). An example of a similarity search originating from research conducted in *Chapter four*, based on molecular structure (chemical descriptor) and on activity profile (biological descriptor) is shown in **Figure 1**.

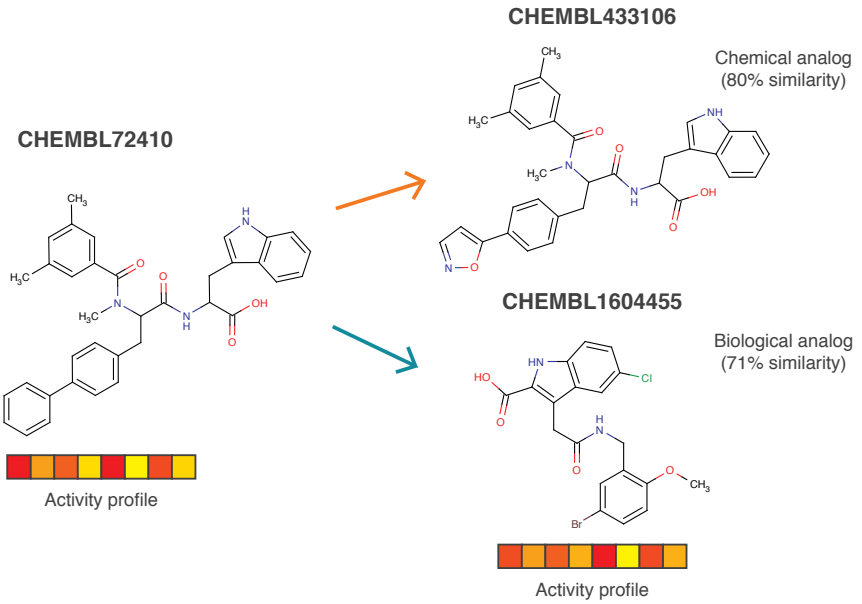


Figure 1. An example of a similarity search originating from research conducted in *Chapter four*, based on molecular structure (chemical descriptor – above) and on activity profile (biological descriptor – below). Chemical analogs often have the same or a very similar scaffold[17] (i.e., molecular framework), whereas biological analogs have similar activity profiles, but may have very different scaffolds. The ability to identify compounds with similar activity profiles, but with different scaffolds (scaffold hopping) is one of the key strengths of biological descriptors.

Chemical analogs are structurally similar to the query compound, with often either the same or a very similar scaffold. By contrast, biological analogs are similar with respect to activity patterns across protein and/or cellular targets, and may be structurally dissimilar to the query compound, enabling scaffold hopping. Recent studies have shown the use of the biological descriptor "high-throughput screening fingerprint" (HTS-FP)[16] to be highly effective at identifying diverse sets of actives[2,16,18] and for mode-of-action analyses.[15,19,20] After the molecules of interest including a set of known actives and inactives (training set) and a set of compounds with unknown actives and inactives (test set) are mapped in space according to a defined descriptor, the next step is to employ a method to identify active compounds in the test set. This method constructs decision boundaries in descriptor space, allowing for the classification of actives and inactives. Here, simple methods such as similarity searching approaches can be used, where only compounds classified as neighbors of known actives given a similarity cutoff are predicted as active.

More sophisticated machine learning classification methods, such as Random Forest[21] and Support Vector Machines[22] (non-linearly) re-scale descriptor space prior to constructing decision boundaries (which are dependent on the cutoff used). This often makes them more suitable for capturing the non-linear relationships between descriptor space and activity commonly found in drug discovery. Other classifiers, such as the Naïve Bayesian classifier[23-25] do not always capture non-linear relationships very well, but can perform well on complex data regardless[26] and require little computational resources. **Figure 2** illustrates how classification methods differ from similarity searching methods with respect to the decision boundaries they construct.

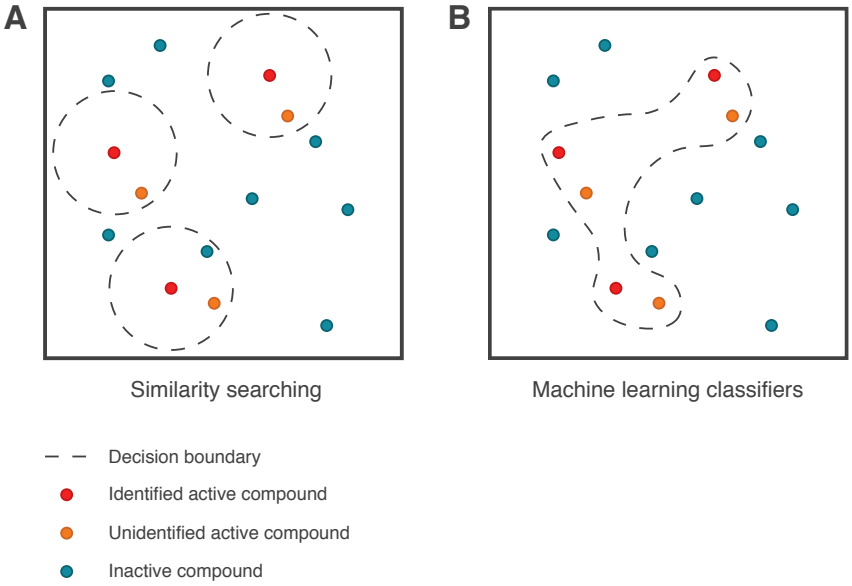


Figure 2. Hypothetical decision boundaries constructed by similarity searching (A) versus those constructed using classification methods (B). Compounds present in the area within the decision boundaries are classified as active according to the method used. Similarity searching assumes that instances within a certain distance from the identified active compounds (red) are active, due to their perceived similarity in descriptor space. Sometimes, inactive compounds are incorrectly classified as active due to this simple assumption (A – bottom decision boundary). Classification methods are able to learn from both active and inactive compound data and construct more accurate decision boundaries. This can improve the classification accuracy.

Similarity searching approaches construct simple decision boundaries, based on the assumption that instances within a certain distance from identified active compounds (red, **Figure 2**) are active, due to their perceived similarity

in descriptor space. This assumption can sometimes lead to false positives (**Figure 2A** – bottom decision boundary). Classification methods use more sophisticated approaches to learn from active and inactive compound data, and are therefore able to construct more accurate decision boundaries, often leading to improved classification accuracy. However, depending on the size, complexity and quality of the data available for training in addition to the classification method used, this advantage can be offset by a high requirement of computational power, sometimes warranting the use of similarity searching. In this thesis, I used similarity searching, the Random Forest[21] classifier and the Naïve Bayesian classifier.[23-25]

Upon classification of molecules as actives or inactives, one has to assess the performance of the method. The receiver operating characteristic (ROC) curve illustrates the performance of a binary classifier as the cutoff defining the decision boundary varies (**Figure 3** – left).[27]

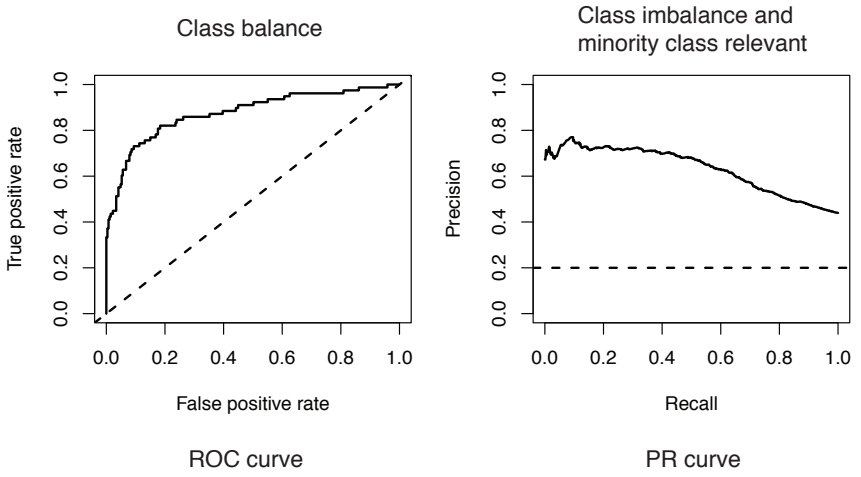


Figure 3. The receiver operating characteristic (ROC) curve (left) and the precision-recall (PR) curve (right). The area under these curves (AUC) is a metric for classifier performance across a range of cutoffs, resulting in a number between zero and one. The dashed lines in both plots represent random classification given 20% of all compounds in the test set are actives. This random classification results in an ROCAUC of 0.5 and a PRAUC corresponding to the relative prevalence of the active class (0.20). The ROCAUC is an appropriate metric when the test set is class-balanced and early enrichment is not required. When early enrichment is desirable – as is often the case in early drug discovery – the Boltzmann-enhanced discrimination of ROC (BEDROC),[27,28] which computes the AUC by weighing early retrieval of active compounds more heavily than later retrieval, is a more appropriate metric.

Here, the true positive rate (TPR) is shown as a function of the false positive rate (FPR). The TPR represents the fraction of true positives (active compounds) retrieved by the classifier, and the FPR is defined as the number of false positives (inactive compounds incorrectly classified as active) divided by the number of true negatives (inactive compounds). The area under the ROC curve (ROCAUC) is used to summarize the performance of the classifier across the range of cutoffs, resulting in a number between zero and one, with 0.5 corresponding to random classification. Therefore, for a classifier to be useful, the ROCAUC must be greater than 0.5 and preferably higher than 0.8 for good classification. While the ROCAUC is commonly used for assessing the performance of a binary classifier (classification of "active versus inactive"), this metric has some drawbacks.

Firstly, the ROCAUC is not an appropriate metric for highly imbalanced datasets, with the number of inactives far outweighing the number of actives[29] as is often the case with bioactivity modeling research in drug discovery. Secondly, in settings where only a relatively small number of compounds need to be selected from a much larger collection (due to limited amounts of resources for testing) the ROCAUC falls short. The reason for this is because it does not place sufficient emphasis on retrieving as many actives as possible in the top most segment of compounds ordered by decreasing likelihood of being active (early enrichment).[27] In case of data imbalance, the area under the precision-recall curve (PRAUC)[29] is a better metric than the ROCAUC, as it captures the effect of the large number of inactive compounds on the model's performance (**Figure 3** – right).[30] When early enrichment is essential, the Boltzmann-enhanced discrimination of ROC (BEDROC) can be used,[27,28] which favors early retrieval of actives.

When only the performance at a specific cutoff is relevant, metrics such as recall (the number of actives retrieved divided by the total number of actives), precision (the number of actives retrieved divided by the total number of compounds classified as active) and F-measure (harmonic mean of recall and precision) can be used. In this thesis, I used the ROCAUC, PRAUC and the BEDROC for bioactivity modeling on HTS data, and the recall, precision and the F-measure for a case study on predicting the mode-of-action of anti-malarial compounds.

Another key topic touched upon in this thesis is the concept of applicability domain (**Figure 4**).

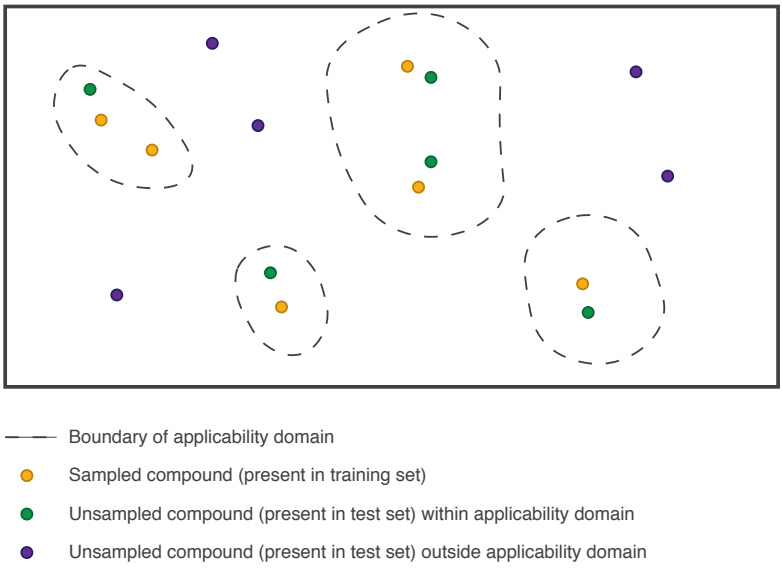


Figure 4. Applicability domain of a classification model. Sampled compounds (yellow) allow the classifier to understand the neighboring descriptor space, and test compounds in that space (green) fall within the applicability domain and can in theory be predicted with more confidence. Areas that are undersampled fall outside the applicability domain, and test compounds in these areas (purple) may not be predicted with certainty.

The quality of machine learning classifiers is directly dependent on the quality and diversity of the underlying training data. Test compounds with close neighbors in the training set (green) are predicted with more confidence regardless of whether they are predicted to be active or inactive, and fall within the applicability domain of the classifier. By contrast, test compounds with no close neighbors in the training set (purple) are more likely to be predicted with less confidence, and possibly lie outside the applicability domain. Awareness of the applicability domain is important, as it allows the user to understand the strengths and limitations of the classifier used and make decisions based on this insight. In *Chapter five* of this thesis, I used a method to systematically sample uncertain areas of descriptor space to design a training set with improved predictivity compared to randomly selected training sets, by aiming to expand the applicability domain.

Objectives of this thesis

In this thesis, I set out to investigate various aspects of bioactivity modeling with the ultimate goal of increasing the efficiency of early-stage screening campaigns (HTS or secondary screening) by anticipating compound activity *in silico*.

Chapter two is a literature review on data-driven approaches used for library design, hit triage and activity modeling in HTS. In particular, the recent rapid progress in bioactivity modeling following the study of Petrone *et al.*[16] is discussed in detail, outlining its significance in the field.

In *Chapter three*, I inspect the relevance of chemical space for bioactivity modeling. Effective model development requires that chemicals be described in terms of a set of characteristics (descriptor) that computers can easily use to assess similarity between molecules. In this chapter, commonly used descriptors are employed to evaluate the diversity of a number of compound sets ranging in size, diversity and origin (e.g., compounds from diversity-oriented synthesis projects, metabolites, drug-like compounds). The degree to which the descriptors used in this study correlated in their assessment of diversity is examined in detail. This provides insight into how the choice of descriptor used affects the sampling of diverse compounds from a large set.

Chapter four illustrates the application of a computational method geared toward systematic compound prioritization. Here, I retrospectively validate the concept of iterative screening on Novartis HTS data. The screening strategy consists of the iterative selection of compounds chemically and biologically similar to actives identified in multiple rounds of testing, leading to consistent increases in efficiency over conventional HTS campaigns.

In *Chapter five*, a data-driven approach is used to derive an "informer compound set". Once screened, this set provides the most information on which yet untested compounds from the remainder of the compound collection to screen next, irrespective of biological target. The derivation of this informer set involves the concept of *active learning*, which attempts to enhance the applicability domain of the classifier. Retrospective validation of this method is performed on public HTS data.[31]

The final research chapter, *Chapter six*, represents a case study in a different context. Here, the application of two machine learning methods (Bayesian target prediction and proteochemometrics modeling) is illustrated for simultaneous polypharmacology and affinity predictions. This approach is used to elucidate the mode-of-action of a collection of anti-malarial

compounds identified in phenotypic screens by GSK, in an attempt to combat neglected diseases.[32]

Finally, *Chapter seven* draws general conclusions from this thesis and provides some future perspectives where I discuss some of my views on (early) drug discovery in academia and the pharmaceutical industry.

References

(1) Cragg, G. M., Newman, D. J., and Snader, K. M. (1997) Natural Products in Drug Discovery and Development. *J. Nat. Prod. 60*, 52–60.

(2) 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.

(3) Drews, J. (2000) Drug Discovery: A Historical Perspective. *Science (80-.). 287*, 1960–1964.

(4) Phatak, S. S., Stephan, C. C., and Cavasotto, C. N. (2009) High-throughput and in silico screenings in drug discovery. *Expert. Opin. Drug Discov. 4*, 947–959.

(5) Wright, L., Wegener, A.-L., and Brierley, C. (2008) Open access to large-scale drug discovery data.

(6) 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.

(7) Bender, A., and Glen, R. C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem. 2*, 3204–3218.

(8) Guha, R., and Van Drie, J. H. (2008) Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model. 48*, 646–658.

(9) Koutsoukas, A., Paricharak, S., Galloway, W. R. J. D., Spring, D. R., IJzerman, A. P., Glen, R. C., Marcus, D., and Bender, A. (2014) How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model. 54*, 230–242.

(10) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model. 50*, 742–754.

(11) Rush, T. S. 3rd, Grant, J. A., Mosyak, L., and Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem. 48*, 1489–1495.

(12) Sauer, W. H., and Schwarz, M. K. (2003) Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci. 43*, 987–1003.

(13) Gillet, V. J., Willett, P., and Bradshaw, J. (1997) Identification of biological activity profiles using substructural analysis and genetic algorithm. *J. Chem. Inf. Comput. Sci. 38*, 165–179.

(14) Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005) Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. 102*, 261–266.

(15) Dančík, V., Carrel, H., Bodycombe, N. E., Seiler, K. P., Fomina-Yadlin, D., Kubicek, S. T., Hartwell, K., Shamji, A. F., Wagner, B. K., and Clemons, P. A. (2014) Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screen. 19*, 771–781.

(16) 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.

(17) Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem. 39*, 2887–2893.
(18) 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.
(19) 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.
(20) Cabrera, A. C., Lucena-Agell, D., Redondo-Horcajo, M., Barasoain, I., Diaz, F., Fasching, B., and Petrone, P. (2016) Compound biological signatures facilitate phenotypic screening and target elucidation. *bioRxiv* 1–27.
(21) Breiman, L. (2001) Random Forests. *Mach. Learn.* 45, 5–32.
(22) Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol. 4*, e1000173.
(23) Nidhi, Glick, M., Davies, J. W., and Jenkins, J. L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model. 46*, 1124–1133.
(24) Nigsch, F., Bender, A., Jenkins, J. L., and Mitchell, J. B. O. (2008) Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model. 48*, 2313–2325.
(25) Paricharak, S., Cortés-Ciriano, I., IJzerman, A. P., Malliavin, T. E., and Bender, A. (2015) Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules. *J. Cheminform. 7*, 15–25.
(26) 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.
(27) Truchon, J., and Bayly, C. I. (2007) Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model. 47*, 488–508.
(28) Riniker, S., and Landrum, G. A. (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform. 5*, 26–42.
(29) Davis, J., and Goadrich, M. (2006) The Relationship Between Precision-Recall and ROC Curves, in *Proceedings of the 23rd International Conference on Machine learning*, pp 233–240.
(30) 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*
(31) 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.
(32) 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.

Chapter two

Data-driven Approaches Used for Compound Library Design, Hit Triage and Bioactivity Modeling in High-throughput Screening (manuscript in preparation)

Shardul Paricharak, Oscar Méndez-Lucio, Aakash Chavan Ravindranath, Andreas Bender, Adriaan P. IJzerman, and Gerard J. P. van Westen

Abstract

High-throughput screening campaigns are routinely performed in pharmaceutical companies to explore activity profiles of chemical libraries for the identification of promising candidates for further investigation. With the aim of improving hit rates in these campaigns, data-driven approaches have been employed to design relevant compound screening collections, enable effective hit triage, and perform activity modeling for compound prioritization. Remarkable progress has been made in the activity modeling area since the recent introduction of large-scale bioactivity-based compound similarity metrics. This is evidenced by increased hit rates in iterative screening strategies and novel insights into compound mode-of-action obtained through activity modeling. Here, we provide an overview of the developments in data-driven approaches, elaborate on novel activity modeling techniques and screening paradigms explored, and outline their significance in high-throughput screening.

Introduction

In the past, knowledge from the areas of pharmacology and medicinal chemistry was combined to design potentially active compounds for testing.[1-3] However, improvements in robotics, automation, and combinatorial chemistry led to the development and increasing use of high-throughput screening (HTS). HTS allowed rapid screening of large compound libraries[3-6] and enabled pharmaceutical companies to explore the bioactivity profiles of compounds covering a larger amount of chemical space[7] with the intention to increase the chances of identifying (diverse) hits for further investigation.

However, multiple non-trivial challenges still exist in HTS. Firstly, the effectiveness in HTS directly depends on the compounds screened, and therefore, the design of compound libraries is of great importance.[8] Secondly, HTS at times cannot be performed for certain assays (such as those involving complex biological systems that do not allow for mass-production), making it an unviable option in such cases.[3,9] Thirdly, measurement errors and artifacts related to assay miniaturization and screening technologies used can complicate the analysis of screening results, making effective triage for follow-up screens a prerequisite for successful campaigns.[8] Lastly, despite improvements in screening technology, HTS campaigns are still costly due to the large amount of resources required in relation to the number of active compounds discovered.[6]

The above-mentioned drawbacks highlight the need for intelligent measures to increase efficiency in HTS. This need, fueled by the increasing amount of bioactivity data available[10] and advances in cheminformatics, has prompted numerous data-driven and computational efforts to improve various aspects of HTS.[11-14]

Approaches suggested for library design include focused design for target classes such as GPCRs or kinases with many known active chemotypes,[2,15,16] and diversity-based design for target classes with few known active chemotypes or for phenotypic assays. For the latter, structural diversity in screening libraries is ensured to increase the chances of finding multiple promising scaffolds for further development across a wide range of assays.[17,18] In addition, much effort has been made to improve hit triage,[19-24] as the selection of actives from primary screens for follow-up screening is not trivial due to the low signal-to-noise ratio in HTS. Finally, virtual HTS (vHTS) approaches are used to prioritize compounds for testing, based on computational model predictions. Recently, ample progress has been made in this area, which we will discuss in detail below.[23,25-31]

In this review, we summarize the recent developments in data-driven applications to improve effectiveness in HTS and discuss the strengths and limitations of these methods. We briefly discuss library design, experimental error management and hit triage. Furthermore, we elaborate on recent developments in bioactivity modeling. Finally, we explore some recently introduced new screening paradigms and highlight their use in further improving efficiency.

Diversity-based library design for targets with few known active chemotypes or phenotypic assays

While over 10^{63} drug-like molecules possibly exist,[32] likely only a fraction of these molecules is therapeutically relevant as evidenced by the success of HTS campaigns comprising "only" 10^6 molecules.[33,34] Therefore, efficient exploration of relevant chemical space is important for targets with few known active chemotypes or phenotypic assays.[35] Diversity-based library design addresses this need by optimizing biological relevance and compound diversity to provide multiple starting points for further development (**Figure 5A**).[17,18]

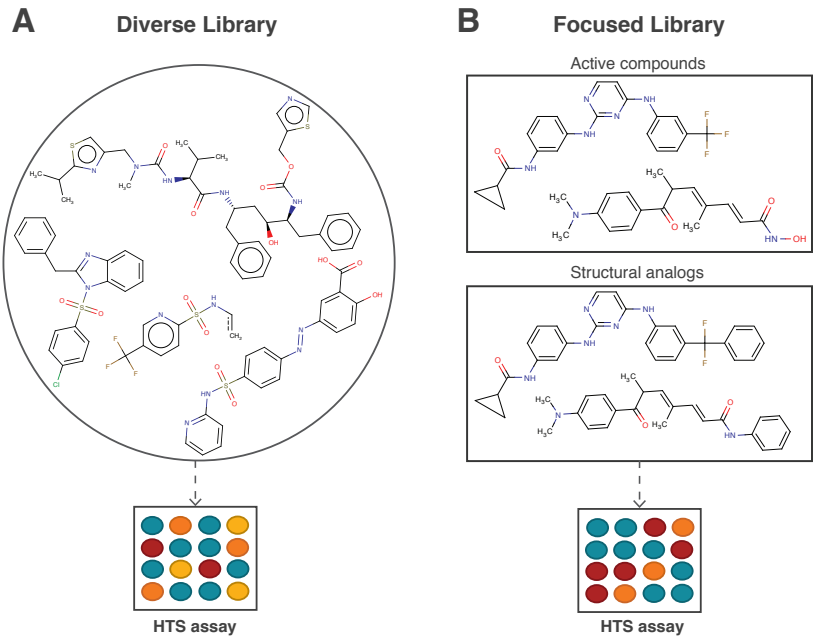


Figure 5. Diverse libraries compared to focused libraries. Structurally diverse libraries are used to efficiently explore relevant chemical space for targets with few known active chemotypes or for phenotypic assays (A).[35] This is performed to provide multiple starting points for further development. Due to the diversity of the compounds tested, a wide range of activities can be observed: from inactive (blue), through somewhat active (yellow) and moderately active (orange), to highly active (red). By contrast, focused libraries are often designed for targets with many known active chemotypes, such as GPCRs, kinases, and in some cases ion channels (B). These libraries focus around active chemotypes found previously, for instance through diversity-based screening.[2,44-46] Here, analogs often exhibit fewer differences in activity.

However, diversity is an ambiguous term,[36,37] as it can be based on a wide range of chemical descriptors (fingerprint-based,[38] shape-based,[39,40] or pharmacophore-based)[41] or even biological descriptors (affinity fingerprints[27,29,42] or HTS-FP),[25] potentially yielding contrasting results.[43] While chemical descriptors characterize compounds in terms of structural and/or physicochemical properties, biological descriptors represent compound phenotypic effects and bioactivity against the druggable proteome. Recent studies at Novartis have shown that these biological descriptors often significantly outperform chemical descriptors regarding hit rate and scaffold

diversity in HTS campaigns, and can even be used in conjunction with chemical descriptors for augmented performance.[13,24,25]

Focused library design for targets with many known active chemotypes

Contrary to diversity-based libraries designed for targets with few known active chemotypes, focused screening libraries are often designed for well-studied targets, such as GPCRs, kinases, and in some cases ion channels. Focused libraries center around active chemotypes found through diversity-based screening (**Figure 5B**)[2,44-46] and can be selected from larger diversity-based libraries using structure-based and/or ligand-based similarity metrics as shown by Tan *et al*.[47] The knowledge of binding mode (such as hinge binding, DFG-out binding, and invariant lysine binding for kinases) is often used during library design to develop ligands with desirable properties.[46] Overall, for target classes with known active chemotypes or with additional information on structure-ligand interaction, focused libraries lead to higher hit rates than diversity-based libraries. This was evidenced in the study by Harris *et al*.[46] where 89% (kinase-focused) and 65% (ion channel-focused) of focused libraries led to an improved hit-rate compared to their diversity-based counterparts. However, despite higher hit rates, focused approaches may not effectively sample diverse chemical space. This could be problematic when certain chemotypes are to be avoided due to off-target effects or intellectual property reasons. Hence, focused libraries are not necessarily a replacement for diversity-based approaches, even for well-studied target classes.

Management of experimental error in HTS

As any experimental technique, HTS is not exempt of experimental errors and the large amount of data obtained from these campaigns make their detection challenging.[48,49] In general, errors in HTS can be classified as random or systematic. Random errors are usually caused by noise and have a low impact in the overall results, as no methodical bias is introduced. By contrast, systematic errors are associated with consistent over- or underestimated activity across the screening collection.[50] Many procedural, technical and environmental reasons exist for systematic errors, such as malfunctioning robots, readout interpretation from plates, reagent evaporation, degradation of target protein, or cell decay.[50,51] Awareness of these problems has prompted

efforts to find new ways of detecting and correcting these errors in order to achieve a better selection of compounds.

Statistics plays an important role in the analysis and detection of errors in HTS.[49,52] Dragiev *et al.*[50] used three statistical approaches to detect systematic errors in HTS data: the χ^2 goodness-of-fit, the Student's *t*–test, and the Discrete Fourier Transform in conjunction with the Kolmogorov-Smirnov test. These methods were used to measure the error in the hit distribution surface, to measure errors for samples with different sizes, and to analyze signal frequency, respectively. In a more recent study, Dragiev *et al.*[51] proposed two widely used methods, namely Matrix Error Amendment (MEA) and Partial Mean Polish (PMP), for correcting errors in HTS with improved results. A deeper discussion of statistical methods for normalization and error correction can be found in two informative reviews.[49,53]

A wide range of software packages[54-58] is available to facilitate analysis and error correction of HTS data (see **Table 1** for an overview).

Table 1. An overview of software available for HTS data analysis. Most software packages enable data analysis and error correction, and more advanced software such as HTS navigator allows for both cheminformatics analysis and visualization.

Software name	**Description**	**Reference (year)**
HTS-Corrector	Analysis and error correction of HTS data	[54] (2006)
HDAT	Web-based HTS data analysis	[55] (2013)
HCS-analyzer	Analysis and error correction of high-content screening data	[56] (2012)
HTS navigator	Cheminformatics analysis, visualization and error correction of HTS data	[57] (2014)
WebFlow	Analysis of HTS cytometry data	[58] (2009)

Earlier programs such as HTS-corrector[54] enable the analysis of background signals, data normalization, and clustering. Building on this foundation more recent and advanced software such as HTS navigator[57] provides features such as loading multiple datasets, visualization, and cheminformatics analysis. The key benefit is that the user can perform a larger part of the analysis on a single platform.

The importance of hit triage

The goal of HTS triage is to prioritize a subset of the large number of detected actives in the primary screen for further investigation and optimization.[8] However, the analysis of HTS data can be complicated by large library sizes and experimental errors caused by artifacts related to assay miniaturization or screening technologies used. A number of filters such as rapid elimination of swill, pan-assay interference compounds (PAINS), the rule of three, and the rule of five are routinely used to discard compounds with undesirable properties (e.g., promiscuity, poor physicochemical properties or presence of problematic functional groups).[8,59-62] While ideally this should take place at the library design stage, analysis of historical HTS data requires that this filtering is applied at the triage stage as well, as often historical assays contain undesirable compounds due to improper filtering at the time of design. This is followed by the selection of diverse sets of actives for follow-up testing based on potency and scaffold structure-activity relationships (SAR).[8,62,63]

Chemically diverse compound sets are preferred over sets comprising many analogs, as the former allows multiple starting points for compound optimization, increasing the overall chances of success. Nevertheless, some analogs in the screening set are desired to enable SAR analysis. HTS data is used to develop models for each chemical class (i.e., scaffold), and active classes are identified based on the relative prevalence of (primary) hits within the class. Actives belonging to an active class are prioritized over those belonging to poorly performing classes, as the latter may more likely be false positives. Additionally, rescuing false negatives is also important; a number of data mining approaches have been explored to this end.[64] Often SAR analysis takes place after secondary screens and concentration-response curves have been performed on a much smaller set of selected compounds. However, a study by Varin *et al*.[63] demonstrated the benefit of including this analysis immediately after the primary HTS screen. Here, primary screening data was preferred over secondary data due to its size and completeness, despite the lower quality. Hit triage results are commonly organized in a scaffold tree with well-defined chemical entities, allowing for intuitive classification and decision-making from a medicinal chemist's point of view.[65]

Developments in virtual HTS (vHTS) and new screening paradigms

vHTS is used in parallel to intelligent library design, error management, and hit triage. vHTS attempts to learn from existing biochemical or phenotypic data and prioritizes subsets of much larger screening libraries for experimental testing.

The wide range of techniques used in vHTS can mainly be divided into two groups: structure-based and ligand-based vHTS. The former relies on three-dimensional structural information (X-ray crystal or NMR structure) of the target protein to study possible interactions with compounds in the screening library.[66,67] The most common structure-based method is molecular docking, which predicts a binding pose for the compound and assigns a score based on the interactions formed in the protein-ligand complex, representing the suitability for experimental testing. By contrast, ligand-based approaches exploit structural information of known active compounds to identify new actives. A number of ligand-based approaches exist: pharmacophore modeling,[68,69] quantitative structure-activity relationship (QSAR) modeling,[70] and similarity searching[71] among others.[66,67]

The low cost and resources required for vHTS combined with the introduction of large public bioactivity databases[10] facilitate its application to many drug discovery campaigns. This has resulted in numerous success stories: the discovery of inhibitors/ligands of DNA methyltransferases (DNMTs),[72,73] kinases[74,75] and GPCRs[76,77] among other relevant targets (see **Table 2** for an overview).[78,79] Nevertheless, the success of vHTS depends on initial data quality and validation procedures.

Table 2. Successful applications of vHTS. Additional examples have been reviewed by Matter and Sotriffer.[87]

Target	Main contribution	Method	Reference (year)
DNMT	Olsalazine, an anti-inflammatory drug as DNMT inhibitor	ligand-based	[72] (2014)
DNMT	Nanaomycin as selective DNMT3b inhibitor	structure-based	[73] (2010)
Chk-1 kinase	Thirty-six inhibitors with IC_{50} values between 68 nM and 110 μM	ligand-based, pharmacophore-based and structure-based	[74] (2003)
JAK3	Identification of a diazaindazole	ligand-based and	[75] (2011)

	scaffold (IC_{50} = 98 nM)	structure-based	
NPY5 receptor	Eleven antagonists ($IC_{50} \leq 1\mu M$)	ligand-based and pharmacophore-based	[76] (2005)
Adenosine receptors	Six high affinity adenosine receptor ligands	ligand-based and binding pocket–based	[77] (2012)
Neurokinin-1 receptor	One compound with IC_{50} = 0.25 μM	pharmacophore-based and structure-based	[78] (2004)
mGlu4 receptor	Six agonists from a library of 720,000 compounds	structure-based	[79] (2005)

With the recent advent of the "high-throughput screening fingerprint" (HTS-FP), which describes compound bioactivity across ~200 biochemical and cell-based assays at Novartis,[25] the concept of bioactivity-based similarity was taken to an unparalleled level. HTS-FP builds on the idea of affinity fingerprints,[27,29,80] allowing a bioactivity-based comparison of compounds. Petrone *et al*.[25] demonstrated the benefit of this descriptor over state of the art chemical descriptors in vHTS and scaffold hopping. This study formed the basis for a body of work on using bioactivity-based similarity searching for mode-of-action analyses[24,26,81,82] and bioactivity modeling, resulting in enhanced (scaffold) hit rates[3,23,24,83] (**Figure 6**). Building on this success, a public version of HTS-FP was later designed based on PubChem bioactivity data.[84]

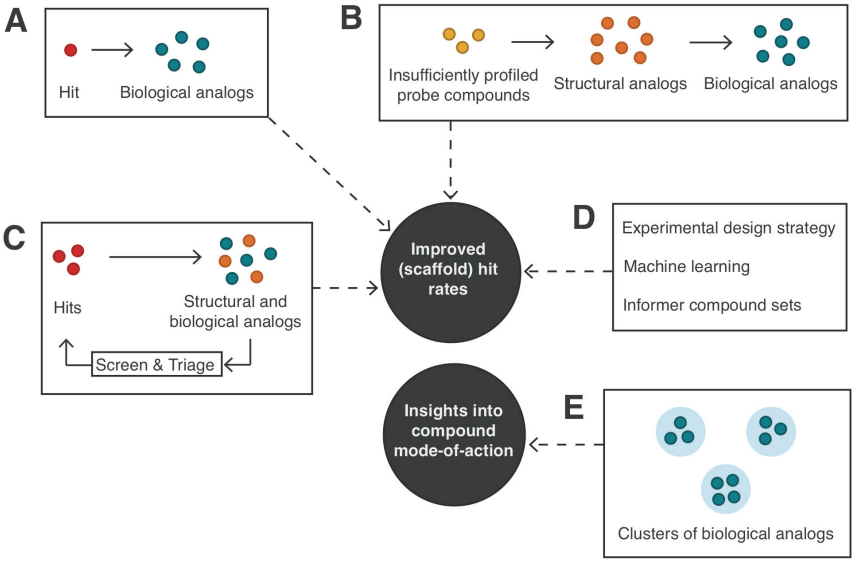


Figure 6. Overview of recent studies improving (scaffold) hit rates and providing insights into compound mode-of-action. Describing compound bioactivity across ~200 assays at Novartis, Petrone *et al.*[25] took the concept of bioactivity-based similarity to an unparalleled level. Here, biological analogs of hits were prioritized for testing (A). Later studies leveraged bioactivity profiles of structural analogs of poorly characterized compounds to select subsets of compounds for virtual screening (B),[24] or employed a screening strategy using biological and chemical similarity metrics in parallel to iteratively expand around hits from multiple rounds of screening (C).[3] Further improvements resulted from changes in experimental design strategy,[83] machine learning methods for predicting actives,[23] and informer sets for routine exploratory screening (D).[86] Other studies used bioactivity-based similarity searching for mode-of-action analyses both at Novartis,[81] Roche,[82] and in the public domain (E).[26]

Wasserman *et al.*[24] developed a method named "bioturbo similarity searching". For insufficiently profiled probe compounds, bioactivity profiles of structural analogs were leveraged to select subsets of compounds for virtual screening. Screening these subsets led to higher (scaffold) hit rates compared to when only structural similarity metrics for expansion around probe compounds were used. Further work addressed the use of bioactivity-based similarity searching for target prediction,[26,81] detection of frequent hitters,[26,62] and iterative selection of activity-enriched subsets of the compound collection for screening.[3] Driven by the gained momentum in machine learning,[85] a comprehensive benchmarking of machine learning classifiers in conjunction with chemical and biological descriptors was performed, with the

overall net result that fusing both HTS-FP and chemical descriptors led to the best performance.[23] Moreover, a study by Paricharak *et al.*[86] described the implementation of an active learning approach to derive "informer compound sets" smaller than 10% of the entire screening collection. Such sets were shown to provide improved predictivity over the remainder of the screening collection compared to randomly selected training sets. Hence the availability of these sets enables routine exploratory screening in an assay-agnostic manner for improved hit expansion.[86]

In pursuit of increased efficiency over conventional HTS campaigns, new screening paradigms have recently been suggested.[3,83] Paricharak *et al.*[3] performed a large-scale validation of iterative screening based on Novartis HTS data. Herein biological and chemical similarity metrics were used in parallel to iteratively expand around hits from multiple rounds of screening, resulting in significantly improved efficiency. Overall, screening 1% of the entire screening collection led to the retrieval of 7500 hits and a cumulative active scaffold coverage of 40%, with efficiency gains realized across a wide range of assay biology.[3] Maciejewski *et al.*[83] suggested an experimental design strategy depending on assay throughput and objective (e.g., hit retrieval or exploration of chemical space for model building). For systems allowing high throughput, conventional expansion around hits was suggested. By contrast, an active learning approach was considered best for iterative screening with smaller compound sets with the explicit aim of developing a model for later use. Here, active learning was preferred due to better sampling of chemical space. When the objective was to optimize cumulative (scaffold) hit rates in iterative screening, the "weak reinforcement strategy" was suggested, where expansion around hits and exploration in under-sampled areas of chemical space was performed simultaneously.[83]

Conclusions

Although HTS has greatly gained momentum over the past decades, much profit can be realized by employing intelligent measures to improve efficiency at the library design, hit triage, and activity modeling stages. Data-driven approaches have consistently been used for improving these aspects, with the aim of systematically prioritizing structurally diverse sets of compounds for further interrogation. HTS-FP and the concept of bioactivity-based similarity have formed the basis for numerous studies showing remarkable improvements in hit retrieval and mode-of-action analyses. However, we note

that while previous work has described harnessing the accumulated knowledge across ~200 assays, the in-depth analysis of activity correlations across independent biochemical and cell-based assays represents an unexplored opportunity. We propose this analysis as an outlook for further investigation, potentially leading to unmapped insights into bioactivity-based similarities between proteins.

References

(1) Drews, J. (2000) Drug Discovery: A Historical Perspective. *Science (80-.). 287*, 1960–1964.

(2) Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today 11*, 277–279.

(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) Mayr, L. M., and Fuerst, P. (2008) The future of high-throughput screening. *J. Biomol. Screen. 13*, 443–448.

(5) Mayr, L. M., and Bojanic, D. (2009) Novel trends in high-throughput screening. *Curr. Opin. Pharmacol. 9*, 580–588.

(6) Phatak, S. S., Stephan, C. C., and Cavasotto, C. N. (2009) High-throughput and in silico screenings in drug discovery. *Expert. Opin. Drug Discov. 4*, 947–959.

(7) Pereira, D. A., and Williams, J. A. (2007) Origin and evolution of high throughput screening. *Br. J. Pharmacol. 152*, 53–61.

(8) Dahlin, J. L., and Walters, M. A. (2014) The essential roles of chemistry in high-throughput screening triage. *Futur. Med. Chem. 6*, 1265–1290.

(9) Astashkina, A., Mann, B., and Grainger, D. W. (2012) A critical evaluation of in vitro cell culture models for high-throughput drug screening and toxicity. *Pharmacol. Ther. 134*, 82–106.

(10) 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.

(11) Huggins, D. J., Venkitaraman, A. R., and Spring, D. R. (2011) Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Biol. 6*, 208–217.

(12) Perez, J. J. (2005) Managing molecular diversity. *Chem. Soc. Rev. 34*, 143–152.

(13) Petrone, P. M., Wassermann, A. M., Lounkine, E., Kutchukian, P., Simms, B., Jenkins, J., Selzer, P., and Glick, M. (2013) Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discov. Today 18*, 674–680.

(14) Willett, P. (1999) Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds. *J. Comput. Biol. 6*, 447–457.

(15) Balakin, K. V., and Bovina, E. V. (2009) Chemogenomics-based design of GPCR-targeted libraries using data-mining techniques, in *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery* (Balakin, K. V., and Ekins, S., Eds.), pp 175–204. Wiley.

(16) Webb, T. R., Venegas, R. E., Wang, J., and Deschenes, A. (2008) Generation of new synthetic scaffolds using framework libraries selected and refined via medicinal chemist synthetic expertise. *J. Chem. Inf. Model. 48*, 882–888.

(17) Shelat, A. A., and Guy, R. K. (2007) Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol. 3*, 442–446.

(18) Fitzgerald, S. H., Sabat, M., and Geysen, H. M. (2006) Diversity Space and Its Application to Library Selection and Design. *J. Chem. Inf. Model. 46*, 1588–1597.

(19) Che, J., King, F. J., Zhou, B., and Zhou, Y. (2012) Chemical and Biological Properties of Frequent Screening Hits. *J. Chem. Inf. Model. 52*, 913–926.
(20) Stanton, D. T., Morris, T. W., Roychoudhury, S., and Parker, C. N. (1999) Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci. 39*, 21–27.
(21) Crisman, T. J., Jenkins, J. L., Parker, C. N., Hill, W. A. G., Bender, A., Deng, Z., Nettles, J. H., Davies, J. W., and Glick, M. (2007) "Plate cherry picking": a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J. Biomol. Screen. 12*, 320–327.
(22) Boyle, N. M. O., Bostro, J., Sayle, R. A., and Gill, A. (2014) Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem. 57*, 2704–2713.
(23) 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.
(24) 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.
(25) 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.
(26) Dančík, V., Carrel, H., Bodycombe, N. E., Seiler, K. P., Fomina-Yadlin, D., Kubicek, S. T., Hartwell, K., Shamji, A. F., Wagner, B. K., and Clemons, P. A. (2014) Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screen. 19*, 771–781.
(27) Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, Å., Bukar, R., Bauer, K. E., Dilley, H., and Rocke, D. M. (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol. 2*, 107–118.
(28) Bender, A., Jenkins, J. L., Glick, M., Deng, Z., Nettles, J. H., and Davies, J. W. (2006) "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. 46*, 2445–2456.
(29) Nguyen, H. P., Koutsoukas, A., Mohd Fauzi, F., Drakakis, G., Maciejewski, M., Glen, R. C., and Bender, A. (2013) Diversity Selection of Compounds Based on "Protein Affinity Fingerprints" Improves Sampling of Bioactive Chemical Space. *Chem. Biol. Drug Des. 82*, 252–266.
(30) Givehchi, A., Bender, A., and Glen, R. C. (2006) Analysis of activity space by fragment fingerprints, 2D descriptors, and multitarget dependent transformation of 2D descriptors. *J. Chem. Inf. Model. 46*, 1078–1083.
(31) 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.
(32) Bohacek, R. S., McMartin, C., and Guida, W. C. (1996) The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev. 16*, 3–50.
(33) Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J., and Shoichet, B. K. (2010) Quantifying Biogenic Bias in Screening Libraries. *Nat. Chem. Biol. 5*, 479–483.
(34) Fox, S., Farr-Jones, S., Sopchak, L., Boggs, A., and Comley, J. (2004) High-throughput screening: searching for higher productivity. *J. Biomol. Screen. 9*, 354–358.
(35) Lipinski, C., and Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature 432*, 855–861.

(36) Koutsoukas, A., Paricharak, S., Galloway, W. R. J. D., Spring, D. R., IJzerman, A. P., Glen, R. C., Marcus, D., and Bender, A. (2014) How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model. 54*, 230–242.
(37) Roth, H. J. (2005) There is no such thing as "diversity"! *Curr. Opin. Chem. Biol. 9*, 293–295.
(38) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model. 50*, 742–754.
(39) Rush, T. S. 3rd, Grant, J. A., Mosyak, L., and Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem. 48*, 1489–1495.
(40) Sauer, W. H., and Schwarz, M. K. (2003) Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci. 43*, 987–1003.
(41) McGregor, M. J., and Muskal, S. M. (1999) Pharmacophore fingerprinting 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci. 39*, 569–574.
(42) Gillet, V. J., Willett, P., and Bradshaw, J. (1997) Identification of biological activity profiles using substructural analysis and genetic algorithm. *J. Chem. Inf. Comput. Sci. 38*, 165–179.
(43) Akella, L. B., and DeCaprio, D. (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol. 14*, 325–330.
(44) Zhang, J., Yang, P. L., and Gray, N. S. (2009) Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer 9*, 28–39.
(45) Van Ahsen, O., and Bomer, U. (2005) High-throughput screening for kinase inhibitors. *Chembiochem. 6*, 481–490.
(46) Harris, C. J., Hill, R. D., Sheppard, D. W., Slater, M. J., and Stouten, P. F. W. (2011) The Design and Application of Target-Focused Compound Libraries. *Comb. Chem. High Throughput Screen. 14*, 521–531.
(47) Tan, L., Lounkine, E., and Bajorath, J. (2008) Similarity Searching Using Fingerprints of Molecular Fragments Involved in Protein−Ligand Interactions. *J. Chem. Inf. Model. 48*, 2308–2312.
(48) Kevorkov, D., and Makarenkov, V. (2005) Statistical Analysis of Systematic Errors in High-Throughput Screening. *J. Biomol. Screen. 10*, 557–567.
(49) Goktug, A. N., Chai, S. C., and Chen, T. (2013) Drug Discovery - Data analysis approaches in high throughput screening, in *Drug discovery* (El-Shemy, H. A., Ed.), pp 201–226. InTech.
(50) Dragiev, P., Nadon, R., and Makarenkov, V. (2011) Systematic error detection in experimental high-throughput screening. *BMC Bioinformatics 12*, 25–38.
(51) Dragiev, P., Nadon, R., and Makarenkov, V. (2012) Two effective methods for correcting experimental high-throughput screening data. *Bioinformatics 28*, 1775–1782.
(52) Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., and Nadon, R. (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol. 24*, 167–175.
(53) Caraus, I., Alsuwailem, A. A., Nadon, R., and Makarenkov, V. (2015) Detecting and overcoming systematic bias in high-throughput screening technologies: a comprehensive review of practical issues and methodological solutions. *Brief. Bioinform. 16*, 974–986.
(54) Makarenkov, V., Kevorkov, D., Zentilli, P., Gagarin, A., Malo, N., and Nadon, R. (2006) HTS-Corrector: new application for statistical analysis and correction of experimental data. *Bioinformatics 22*.
(55) Liu, R., Hassan, T., Rallo, R., and Cohen, Y. (2013) HDAT: web-based high-throughput screening data analysis tools. *Comput. Sci. Discov. 6*, 14006–14016.
(56) Ogier, A., and Dorval, T. (2012) HCS-Analyzer: open source software for high-content screening data correction and analysis. *Bioinformatics 28*, 1945–1946.

(57) Fourches, D., Sassano, M. F., Roth, B. L., and Tropsha, A. (2014) HTS navigator: freely accessible cheminformatics software for analyzing high-throughput screening data. *Bioinformatics 30*, 588–589.
(58) Hammer, M. M., Kotecha, N., Irish, J. M., Nolan, G. P., and Krutzik, P. O. (2009) WebFlow: a software package for high-throughput analysis of flow cytometry data. *Assay Drug Dev. Technol. 7*, 44–55.
(59) Walters, W. P., and Namchuk, M. (2003) A guide to drug discovery: designing screens: how to make your hits a hit. *Nat. Rev. Drug. Discov. 2*, 259–266.
(60) Lovering, F., Bikker, J., and Humblet, C. (2009) Escape from Flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem. 52*, 6752–6756.
(61) Walters, W. P. (2012) Going further than Lipinski's rule in drug design. *Expert. Opin. Drug Discov. 7*, 99–107.
(62) Baell, J., and Walters, M. A. (2014) Chemical con artists foil drug discovery. *Nature 513*, 481–483.
(63) Varin, T., Gubler, H., Parker, C. N., Zhang, J., Raman, P., Ertl, P., and Schuffenhauer, A. (2010) Compound Set Enrichment: A novel approach to analysis of primary HTS data. *J. Chem. Inf. Model. 50*, 2067–2078.
(64) Glick, M., Jenkins, J. L., Nettles, J. H., Hitchings, H., and Davies, J. W. (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model. 46*, 193–200.
(65) Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M. A., and Waldmann, H. (2007) The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model. 47*, 47–58.
(66) Lavecchia, A., and Giovanni, C. (2013) Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem. 20*, 2839–2860.
(67) Bielska, E., Lucas, X., Czerwoniec, A., Kasprzak, J. M., Kaminska, K. H., and Bujnicki, J. (2011) Virtual screening strategies in drug design - methods and applications. *BioTechnologia 3*, 249–264.
(68) Leach, A. R., Gillet, V. J., Lewis, R. A., and Taylor, R. (2010) Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem. 53*, 539–558.
(69) Van Drie, J. H. (2011) Generation of three-dimensional pharmacophore models. Wiley Interdiscip. *Rev. Comput. Mol. Sci. 3*, 449–464.
(70) Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. (2014) QSAR modeling: where have you been? Where are you going to? *J. Med. Chem. 57*, 4977–5010.
(71) Stumpfe, D., and Bajorath, J. (2011) Similarity searching. Wiley Interdiscip. *Rev. Comput. Mol. Sci. 1*, 260–282.
(72) Méndez-Lucio, O., Tran, J., Medina-Franco, J. L., Meurice, N., and Muller, M. (2014) Toward Drug Repurposing in Epigenetics: Olsalazine as a Hypomethylating Compound Active in a Cellular Context. *ChemMedChem. 9*, 560–565.
(73) Kuck, D., Singh, N., Lyko, F., and Medina-Franco, J. L. (2010) Novel and selective DNA methyltransferase inhibitors: Docking-based virtual screening and experimental evaluation. *Bioorg. Med. Chem. 18*, 822–829.
(74) Lyne, P. D., Kenny, P. W., Cosgrove, D. A., Deng, C., Zabludoff, S., Wendoloski, J. J., and Ashwell, S. (2004) Identification of Compounds with Nanomolar Binding Affinity for Checkpoint Kinase-1 Using Knowledge-Based Virtual Screening. *J. Med. Chem. 47*, 1962–1968.

(75) Chen, X., Wilson, L. J., Malaviya, R., Argentieri, R. L., and Yang, S.-M. (2008) Virtual Screening to Successfully Identify Novel Janus Kinase 3 Inhibitors: A Sequential Focused Screening Approach. *J. Med. Chem. 51*, 7015–7019.
(76) Guba, W., Neidhart, W., and Nettekoven, M. (2005) Novel and potent NPY5 receptor antagonists derived from virtual screening and iterative parallel chemistry design. *Bioorg. Med. Chem. Lett. 15*, 1599–1603.
(77) 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., and Bender, A. (2012) Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. *J. Med. Chem. 55*, 7010–7020.
(78) Evers, A., and Klebe, G. (2004) Successful Virtual Screening for a Submicromolar Antagonist of the Neurokinin-1 Receptor Based on a Ligand-Supported Homology Model. *J. Med. Chem. 47*, 5381–5392.
(79) Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., and Bertrand, H.-O. (2005) Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem. 48*, 2534–2547.
(80) Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005) Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. 102*, 261–266.
(81) 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.
(82) Cabrera, A. C., Lucena-Agell, D., Redondo-Horcajo, M., Barasoain, I., Diaz, F., Fasching, B., and Petrone, P. (2016) Compound biological signatures facilitate phenotypic screening and target elucidation. *bioRxiv* 1–27.
(83) 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.
(84) Helal, K. Y., Maciejewski, M., Gregori-Puigjané, E., Glick, M., and Wassermann, A. M. (2016) Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J. Chem. Inf. Model. 56*, 390–398.
(85) Reker, D., and Schneider, G. (2015) Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today 20*, 458–465.
(86) 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*
(87) Matter, H., and Sotriffer, C. (2011) Applications and Success Stories in Virtual Screening., in *Virtual Screening*, pp 319–358. Wiley-VCH Verlag GmbH & Co. KGaA.

Chapter three

How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space

Abstract

Chemical diversity is a widely applied approach to select structurally diverse subsets of molecules, often with the objective of maximizing the number of hits in biological screening. While many methods exist in the area, few systematic comparisons using current descriptors in particular with the objective of assessing diversity in bioactivity space have been published, and this shortage is what the current study is aiming to address. In this work, 13 widely used molecular descriptors were compared, including fingerprint-based descriptors (ECFP4, FCFP4, MACCS keys), pharmacophore-based descriptors (TAT, TAD, TGT, TGD, GpiDAPH3), shape-based descriptors (rapid overlay of chemical structures (ROCS) and principal moments of inertia (PMI)), a connectivity-matrix-based descriptor (BCUT), physicochemical-property-based descriptors (prop2D), and a more recently introduced molecular descriptor type (namely, "Bayes Affinity Fingerprints"). We assessed both the similar behavior of the descriptors in assessing the diversity of chemical libraries, and their ability to select compounds from libraries that are diverse in bioactivity space, which is a property of much practical relevance in screening library design. This is particularly evident, given that many future targets to be screened are not known in advance, but that the library should still maximize the likelihood of containing bioactive matter also for future screening campaigns. Overall, our results showed that descriptors based on atom topology (i.e., fingerprint-based descriptors and pharmacophore-based descriptors) correlate well in rank-ordering compounds, both within and between descriptor types. On the other hand, shape-based descriptors such as ROCS and PMI showed weak correlation with the other descriptors utilized in this study, demonstrating significantly different behavior. We then applied eight of the molecular descriptors compared in this study to sample a diverse subset of sample compounds (4%) from an initial population of 2587 compounds, covering the 25 largest human activity classes from ChEMBL and measured the coverage of activity classes by the subsets. Here, it was found that "Bayes Affinity Fingerprints" achieved an average coverage of 92% of activity classes. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes were represented in the selected compounds respectively, followed by BCUT, prop2D, MACCS, and PMI (in order of decreasing performance). In addition, we were able to show that there is no visible correlation between compound diversity in PMI space and

in bioactivity space, despite frequent utilization of PMI plots to this end. To summarize, in this work, we assessed which descriptors select compounds with high coverage of bioactivity space, and can hence be used for diverse compound selection for biological screening. In cases where multiple descriptors are to be used for diversity selection, this work describes which descriptors behave complementarily, and can hence be used jointly to focus on different aspects of diversity in chemical space.

Introduction

Computational methods play a pivotal role in modern drug discovery, extending from de novo computer-aided drug-design methods[1,2] to tasks such as storing and analyzing large combinatorial chemical libraries.[3] The size of chemical space is difficult to quantify, but there is no doubt that it is very large in nature and, according to one estimate, there are ~10^{63} organic small molecules that could be formed of up to 30 heavy atoms.[1,4] Considering the fact that likely only a very small fraction of that space is therapeutically relevant, it is of high importance to develop novel computational approaches that will allow efficient exploration and effective selection of molecules that could be tested for bioactivity against proteins of interest.[5]

One concept in this direction that has gained importance in recent decades, in particular with the advent of modern combinatorial chemistry and high-throughput screening (HTS), is chemical diversity. Chemical diversity, a concept complementary to (but not just the mirror image of) chemical similarity,[6,7] is routinely utilized in library design and compound selection to quantitatively evaluate the presence of distinct structural features (however defined) present in chemical libraries.[8,9] To this end numerous computational methods are currently available which allow for the selection of structural diverse subsets that maximize the chemical space, i.e., selecting a set of compounds with the maximum degree of structural variation, while retaining a manageable number of molecules to be screened or tested for novel activities.[10]

Chemical diversity is by no means a uniquely defined concept, and it has been argued that it could only be measured by relevant external criteria and thus cannot be inherently "objective".[11] Nevertheless, it has developed into a concept of high practical relevance in the field of cheminformatics and, in particular, the design of screening libraries, as it allows one to quantify the similarity (or dissimilarity) of two or more chemical libraries and rationally

select chemically diverse compounds from a much larger population of molecules. This approach is particularly suitable when knowledge about the chemical matter active on a protein target is limited (in this case, more focused compound selection would often be performed), or when the aim is to design a general screening set that can be applied to multiple protein targets.[8]

An important aspect when measuring chemical diversity is the choice of molecular descriptors, which is used as input for distance or similarity measures to quantify this measure. An ideal molecular descriptor might, for example, show good correlation with human perception of chemical diversity, so that it resembles the human mind in decision-making processes. However, this is, in practice, difficult to realize for at least two reasons: first, because it is difficult to say with certainty how humans actually assess chemical structures that are displayed in front of them,[12] and, second, because there is a remarkable inconsistency in assessment, both between chemists, and also when displaying the same structures to a chemist repeatedly.[13] Hence, resembling human perception might actually not be a desirable goal to pursue in the end, and quantitative measures related to the problem at hand might be more suitable to measure the performance of diversity assessment methods. As described in detail later, the quantitative "external" measure that we decided to pursue in the current work was bioactivity space coverage, given that in many cases diversity selection of compounds aims at assembling a library of small molecules with increased chances of identifying hits against both current and future targets.

Recent studies have demonstrated that the use of different descriptors could generate significantly different results when selecting subsets of diverse molecules,[14–16] underlining that the choice of molecular descriptor clearly affects (or biases) the perception of chemical diversity present in a library. Furthermore, as was shown by Fergus et al., the diversity of a library is dependent on the number of library members, and very small libraries could give counterintuitive estimates of diversity and should be treated with caution.[17] Different diversity assessment methods can yield vastly different results,[18] depending on what type of chemical libraries they are used to analyze, the size of the libraries, as well as the source of the molecules, which can be compounds stemming from combinatorial chemistry, or natural products.[19]

Currently, various methods are routinely being used to assess chemical diversity, including fingerprint-based,[16] shape-based,[20] and pharmacophore-based methods.[21] Fingerprint-based methods compare small molecules in terms of the presence or absence of a set of substructural or fingerprint features (derived from molecular graph representations), hereby taking into account atom connectivity, and are widely used in virtual screening.[22] Alternatively, shape-based methods encode molecular conformational information, which can be internal distances or external molecular properties, which are then applied to compare molecules based on those properties. Examples of such shape-based methods are ROCS descriptors, which compare molecules based on their molecular shapes, by assessing atom-centered overlapping Gaussians and calculating the maximal intersection of the volume between molecules.[23] Furthermore, pharmacophore-based methods compare molecular similarity in terms of the presence or absence of pharmacophoric features (which may, in turn, often be represented as fingerprints).[21]

Despite the wide usage of all the above-mentioned methods, each descriptor focuses only on one aspect of the chemical information available. For example, shape-based methods are scaffold-independent, whereas pharmacophoric descriptors focus on pharmacophoric points and do not take into account the entire molecular surface, and structural keys encode only the presence or absence of predefined substructural features but not the connectivity among them. An alternative approach to quantify molecular diversity that has been explored recently (and also in this work) is based on "in silico" bioactivity profiles, which maps chemical structural space into a predicted bioactivity profile against a large number of protein targets.[24] Given that diversity in bioactivity space is often the main aim of diversity selection projects, basing the decision of how diverse compounds are on their bioactivity profiles might well be a purely empirical, but still rather suitable, decision to make.

Common to all of the above descriptors, once chemical structures have been encoded in a computer-accessible way, selection algorithms come into play that, in principle, can be based on any of the descriptors mentioned above.[10] The objective of these methods is to select subsets of compounds with maximum structural diversity from an initial large pool of compounds, while retaining the overall diversity of the initial population of molecules. Because of the large size of libraries, which can be in the order of 10^6 or larger in HTS

campaigns, these are mainly heuristic methods, since exhaustive enumeration of all possible subsets would be computationally unfeasible and can be categorized as[25] (i) maximum dissimilarity-based algorithms,[26] (ii) clustering,[27] (iii) partitioning of cell-based approaches,[28] and (iv) optimization methods.[29,30] All methods provide approaches to cherry pick diverse sets of compounds from large libraries; however, they still depend on molecular descriptors to compare compounds and, therefore, are affected by the shortcomings and the behavior of the descriptor applied.

In this study, we will focus on the descriptor aspect of chemical diversity selection, given that no comparative study using bioactivity coverage as an objective function has been published yet, according to the best of the authors' knowledge, despite its importance for selecting diverse screening subsets. To this end, the behavior of 13 widely employed descriptor types was assessed, which fall into four main categories, namely, fingerprints, shape-based methods, pharmacophoric methods, and two-dimensional (2D) properties. In addition, we utilized "Bayes Affinity Fingerprints" as descriptors,[31,24] that represent molecular structures based on their *in silico* bioactivity profiles.

These descriptors were applied for diversity selection across different chemical libraries varying both in size and diversity, and were assessed with respect to their similarities and differences in rank-ordering compounds in diversity selection procedures by employing the Spearman's rank correlation coefficient. In addition, coverage of bioactivity space was measured to assess descriptor performance, in addition to differences in behavior. Hence, the objective of this study is to assess correlation among widely employed chemical descriptors across a large set of libraries, in order to obtain a better understanding of the situations in which these descriptors correlate and when they do not, as well as to evaluate their ability to cover large numbers of bioactivity classes in the selected subsets. This is of relevance for selecting compound subsets for biological screening, in particular, in cases where either different target families or orphan targets will be screened.

Methods

Molecular Datasets. The behavior of the different diversity assessment and selection approaches was assessed on diverse sets of small molecule libraries, namely, compounds generated via Diversity-Oriented Synthesis (DOS)

approaches,[32] metabolites from HMBD,[33] DrugBank,[34] PubChem,[35] and ChEMBL.[36]

FN, CEH, HEB, DRS, and Da Libraries. All of the libraries from this category stem from Diversity-Oriented Synthesis (DOS) approaches, and they were used since they claim to contain large chemical diversity by their very nature.[37] The FN, CEH, and HEB datasets consist of 45 nonpeptidic macrocyclic compounds in total and are part of a larger library of such compounds generated using a DOS approach.[38] The compounds all contain macrocyclic rings, a structural motif which is argued to be of value for targeting the extended binding interface associated with protein−protein interactions.[39,40] The DRS library consists of 28 compounds, which were not based on any general scaffold type. The Da library consists of 27 small molecules generated using a branching DOS strategy.[41,42] All of these DOS libraries were synthesized with the intent of providing hits against diverse biological targets, and, indeed, the screening of the Da and DRS libraries has identified compounds with antibacterial activity already.[41,42]

HMDB: Human Metabolome Database (HMDB). The Human Metabolome Database[33] is a comprehensive resource of small endogenous molecule metabolites found in the human body. For this study, HMDB version 2.5 was stored locally in SDF format, containing 8535 compounds in total. Random selection was applied, followed by a filtering criterion of MW < 900 Da. In total, 981 molecules were selected for the current study.

DrugBank. DrugBank[34] constitutes a comprehensive resource for drugs and drug target information. For this study, DrugBank version 3 was used and stored locally in SDF format. Compounds with a molecular weight of 900 Da or less were randomly selected, leading to a total of 1036 drugs and druglike molecules considered in this study.

PubChem. PubChem[35] is a large open repository for small molecules and biological properties of small molecules for public access, hosted by the U.S. National Institutes of Health (NIH). For this study, the PubChem FTP service was accessed and 10 random subsets of compounds from the full database were selected and downloaded locally in SDF file format. Consecutive random selection steps were applied to select molecules. In addition, the molecular weight was set not to exceed 950 Da, leading to a total of 947 compounds selected and used in this study. No diversity selection algorithm was applied prior to analyzing the results, to avoid introducing descriptor bias.

ChEMBL. ChEMBL[36] is a database containing binding, functional, and ADMET information for a large number of druglike bioactive molecules maintained by EMBL-EBI. ChEMBL version 14 was utilized for this study, downloaded, and installed on a local MySQL server. The 50 most-populated human protein targets were selected, based on the number of compounds annotated with K_i, IC_{50}, EC_{50} and K_d values equal to or better than 1 μM. These data consisted of various types of targets including enzymes (proteases, lyases, reductases, hydrolases, and kinases) representing 48% of the classes, membrane receptors (GPCRs and non-GPCRs) representing 44% of the classes and transcription factors and transporters each representing 4% of the classes. The target classes contained 1573 compound associations on average, varying from 1014 to 2971 data points. Subsequently, 2587 compounds were randomly selected from the 25 largest classes. This resulted in classes containing 103 data points on average, with the smallest class containing 12 data points and the largest class containing 190 data points. This dataset was utilized to compare the performance of how molecular descriptors sample diverse subsets of compounds and achieve protein target coverage.

TIMBAL. TIMBAL[43] is a database containing small molecules that modulate protein−protein interactions. All compounds with annotated K_i, IC_{50}, EC_{50}, or K_d values of 10 μM or better were selected, which resulted in a total of 1995 unique compounds across 34 target classes after standardization (as described in the following subsection, "Library Preparation").

Library Preparation. Molecules were standardized using ChemAxon's Standardizer with the options Remove salts (keep largest fragment), Neutralize, Remove Explicit Hydrogens, Aromatize, Mesomerize, and Tautomerize.[44] Following standardization, molecules were loaded to Molecular Operating Environment 2011.10[45] (MOE) and three-dimensional (3D) molecular conformations were calculated using MOE,[45] applying the Rebuild 3D option, while retaining existing chirality (default options).

Molecular Descriptors. Twelve (12) widely employed structural molecular descriptors and one descriptor based on predicted bioactivity spectra, namely, the "Bayes Affinity Fingerprints", were utilized in this study for the representation of molecules. Molecular descriptors are listed in **Table 3** and briefly described in the following.

Table 3. Molecular descriptors used in this study and software implementation.

Descriptor type	Descriptor name	Implemented in	Description

Fingerprint-based	ECFP4	MOE v2011.10[42]	atom type, extended connectivity fingerprint, maximum distance = 4
	FCFP4	MOE v2011.10[42]	functional-class-based, extended connectivity fingerprint, maximum distance = 4
	MACCS	MOE v2011.10[42]	166 predefined MDL keys (public set)
Connectivity-matrix-based	BCUT	MOE v2011.10[42]	atomic charges, polarizabilities, H-bond donor and acceptor abilities, and H-bonding modes of intermolecular interaction
Shape-based	rapid overlay of chemical structures (ROCS), combo Tanimoto (shape and electrostatic score)	OpenEye v3.1.2[48]	shape-based molecular similarity method; molecules are described by smooth Gaussian function and pharmacophore points
	PMI	MOE v2011.10[42]	normalized principal moment-of-inertia ratios
Pharmacophore-based	GpiDAPH3	MOE v2011.10[42]	graph-based 3-point pharmacophore, eight atom types computed from three atom properties (in pi system, donor, acceptor)
	TGD	MOE v2011.10[42]	typed graph distances, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
	TAD	MOE v2011.10[42]	typed atom distances, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
	TGT	MOE v2011.10[42]	typed graph triangles, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
	TAT	MOE v2011.10[42]	typed atom triangles, atom typing (donor, acceptor, polar, anion, cation, hydrophobe)
Bioactivity-based	Bayes affinity fingerprints	in-house-developed *in silico* bioactivity prediction model[51]	bioactivity model based on multicategory Bayes classifier trained on data from ChEMBL v. 14
Physicochemical-property-based	prop2D	MOE v2011.10[42]	physicochemical properties (such as molecular weight, atom counts, partial

			charges, hydrophobicity etc.)

(i) Fingerprint-based descriptors:

(1) *MACCS keys (MOE).*[46] 166 predefined substructural key sets of the public subset as implemented in MOE, which were originally designed for quicker database retrieval of compounds with certain predefined chemical functionalities.

(2) *ECFP4 and FCFP4 (MOE).*[47] Circular fingerprints as implemented in MOE, where E stands for atom type and F stands for functional class. Extended connectivity fingerprints are derived from variation of the Morgan algorithm,[48] and this descriptor type has been shown to capture much information relevant to the bioactivity of a compound.[49,50]

(ii) Pharmacophore-based descriptors:

(3) *GpiDAPH3 (MOE).*[51] Graph-based three-point pharmacophore employing any set of three possible atom types, namely, "in pi system", "donor", and "acceptor" atom.

(4) *TAD, TAT, TGD and TGT (MOE).* Typed atom distances (TAD), typed atom triangle (TAT), typed graph distances (TGD), and typed graph triangles (TGT). Six different atom types are possible: donor, acceptor, polar, anion, cation, and hydrophobic.

(iii) Shape-based descriptors:

(5) *ROCS (OpenEye).*[23,52,53] Molecular shapes are described by smooth Gaussian function and pharmacophoric points. "Combo Tanimoto" was used as a similarity function.

(6) *PMI (MOE).*[54] Three principal moments of inertia (PMI) derived from 3D structures as implemented in MOE.

(7) *BCUT (MOE).*[55] Four-dimensional (4D) BCUT_PEOE descriptors as implemented in MOE. BCUT descriptors are based on atomic charges, polarizabilities, H-bond donor, and acceptor abilities and H-bonding modes of intermolecular interaction.

(iv) 2D descriptors:

(8) *prop2D.* The first 10 principal components of all 2D physicochemical properties as implemented in MOE (v2011.10), containing properties such as molecular weight, atom counts, polar surface area, etc.

(v) Bayes Affinity Fingerprints:[56]

In silico predicted bioactivity spectra of small molecules generated using an in-house-developed bioactivity model based on the multicategory Naıv̈e Bayesian classifier and bioactivity data extracted from ChEMBL. Compounds are initially described by circular molecular fingerprints and then are subjected to a target prediction model containing 134450 bioactive compounds that cover 477 human protein targets. Protein targets are then ranked for each compound based on the likelihood of being active.

Distance Metrics.

Euclidean distances. Euclidean distances were employed as a distance function for descriptors that assume continuous values, namely PMI, Bayes affinity fingerprints, ROCS, BCUT and 2D physicochemical descriptors. Euclidean distances were calculated according to eq 1:[57]

$$d(\boldsymbol{p},\boldsymbol{q}) = d(\boldsymbol{q},\boldsymbol{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are the descriptors vectors of two molecules containing $\boldsymbol{n}$ dimensions.

Hamming distance. The Hamming distance was employed as a distance function for descriptors that take binary values (0 or 1), namely circular, structural and pharmacophoric fingerprint based methods (ECFP4, FCFP4, MACCS, TGD, TGD, TAT, TAD, GpiDAPH3). The Hamming distance, which originates from information theory, is calculated for two equal *n*-length vectors as the number of positions at which the corresponding bits are different.[58] The use of the Hamming distance was preferred over the Tanimoto metric in this study, as the Tanimoto metric is a normalized metric taking values between the intervals [0,1], while the Hamming distance is not normalized. For two vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$, the Hamming distance was calculated according to eq 2:

$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n}|p_i - q_i| \quad (2)$$

Spearman's Rank Correlation Coefficient. The Spearman's rank correlation coefficient (ρ), was employed as correlation coefficient for assessing statistical dependence between rankings obtained among molecules as a result of applying different descriptors (each descriptor led to a unique ranking of molecules depending on the properties encoded by the descriptor, which was

then compared across descriptor pairs for each dataset used).[59] Spearman's rank correlation coefficient can assume values in the interval [-1,1], where -1 or 1 indicates perfect positive or negative correlation, and values of zero indicate absence of any correlation between the two variables analyzed. Spearman's rank correlation coefficient was calculated according to eq 3:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (3)$$

where x_i and y_i were the resulted rankings obtained from the distance matrices for each molecular descriptor used.

Principal Component Analysis (PCA). In order to obtain visualization of the arrangement of descriptors in multidimensional space, PCA was performed using the R statistical environment (version 2.15.2)[60] and the three first-principal components were visualized using Vortex.[61]

Assessment of Bioactivity Coverage. The performance of each compound diversity selection method was assessed based on the coverage of activity classes achieved by sampling a 1% diverse subset from the initial population of the compounds and by counting number of activity classes being retrieved. The experiments were repeated three times and the average number of protein targets presented in each sampled set was assessed. The descriptor-based diversity selection was performed utilizing MOE's function "Calculate Diverse Subset" with the option Output limit set to 100. In case of Bayes Affinity Fingerprints the approach as described by Nguyen *et al.*[24] was applied. Not all molecular descriptors were utilized in this step due to high computational cost (e.g., calculating a similarity matrix among compounds in the PubChem dataset based on ROCS descriptors took approximately one week on an Intel Core 2 Duo desktop with 8 GB of RAM). Instead, eight representative descriptors were used, as shown in **Table 4**.

Table 4. Activity classes covered by sampling a diverse subset of 4% (100 compounds) from an initial set of 2587 compounds extracted from the 25 largest ChEMBL human activity classes. Using Bayes Affinity Fingerprints considering only predicted protein targets with a Bayes Score above 30, 92% of the initial activity classes were sampled, hence outperforming all other descriptors marginally. Using the descriptors ECFP4 and GpiDAPH3, TGT and random sampling, 91%, 84%, 84%, and 84% of the activity classes were represented in the selected compounds respectively. Random sampling retrieved approximately 84% of activity classes (averaged over three attempts), showing similar performance on this dataset with GpiDAPH3, TGT and BCUT, while outperforming molecular descriptors such as prop2D (80%), MACCS (80%), and PMI (75%).

Descriptor	**% Activity classes sampled (subset size of 25 classes) averaged over three attempts (relative ranking)**	**Number of activity classes sampled out of 25 classes – 1st attempt**	**Number of activity classes sampled out of 25 classes – 2nd attempt**	**Number of activity classes sampled out of 25 classes – 3rd attempt**	**Average**
Bayes affinity fingerprints ("Cutoff 30")	92% (1st)	22	24	23	23
ECFP4	91% (2nd)	22	23	23	22.7
GpiDAPH3	84% (3rd)	21	22	21	21
TGT	84% (4th)	20	21	22	21
Random Sampling	84% (5th)	21	21	19	21
BCUT	83% (6th)	21	20	21	20.7
prop2D	80% (7th)	21	21	19	20
MACCS	80% (8th)	19	20	20	20
PMI	75% (9th)	19	18	19	18.7

Results and discussion

Results obtained on the overall correlation of descriptors used in diversity assessment and averaged across all libraries are visualized in **Figure 7**.

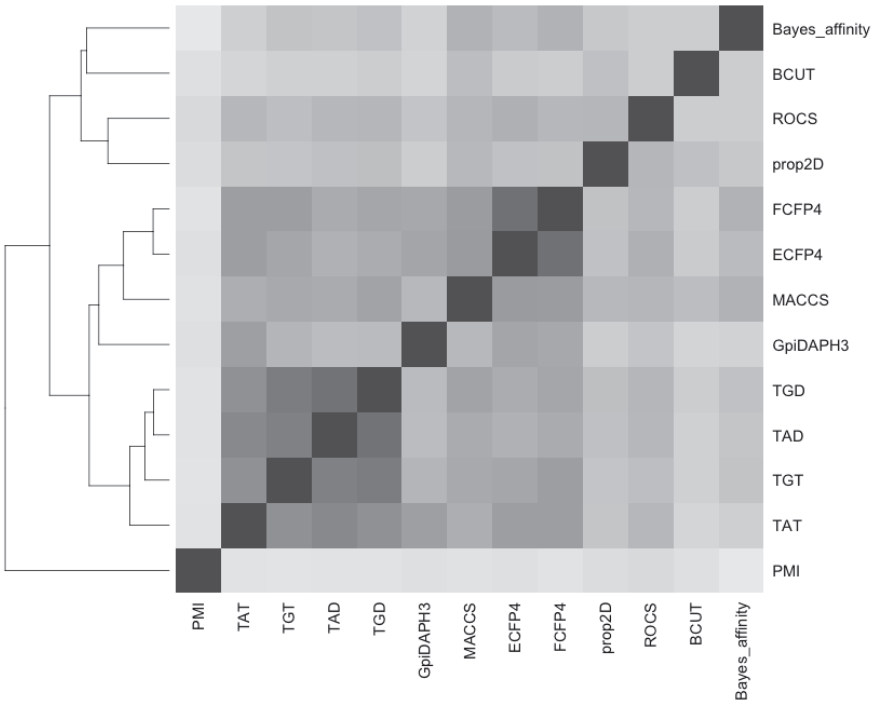


Figure 7. Spearman's rank correlation coefficients obtained from descriptors used in this study averaged over all included libraries. Darker colors indicate higher Spearman's rank correlation, whereas lighter colors indicate lower correlation. The descriptor PMI demonstrated the least correlation with any other descriptor indicating that this type of descriptor behaves significantly differently from any others included here. Overall, it can be seen that pharmacophore-based descriptors such as TAT, TGT, TAD and TGD show some correlation with fingerprint-based descriptors (ECFP4, FCFP4 and MACCS, with MACCS keys also showing some correlation with other descriptor types).

The PCA visualization performed on the matrix obtained from Spearman's rank correlation among molecular descriptors is presented in **Figure 8**, where the first three principal components explained 74% of the accumulative variance as shown in **Figure 9**, whereas ~90% of the variance in descriptor space is explained by the first five principal components.

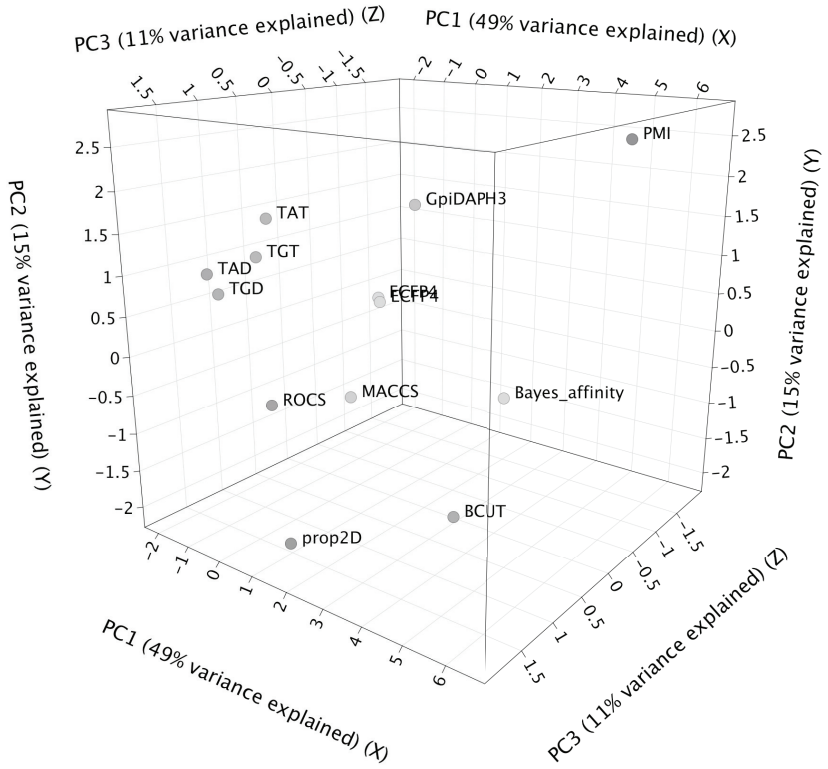
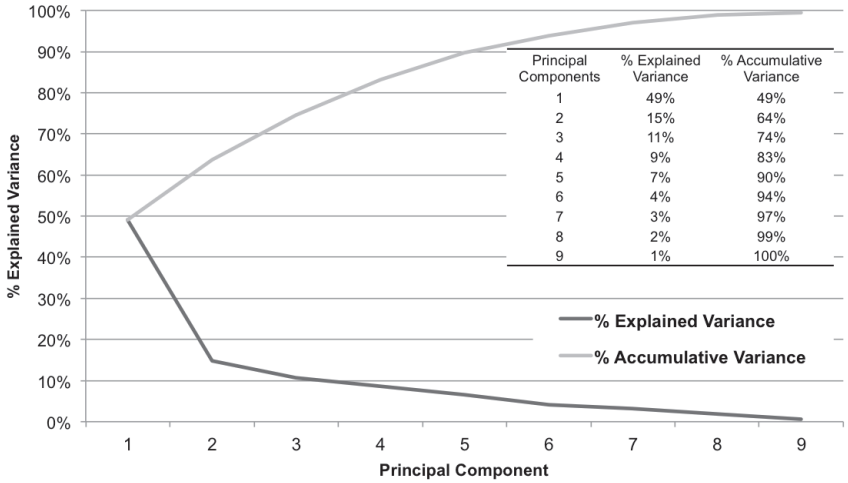


Figure 8. Similarity of molecular descriptors in perceiving the diversity of chemical libraries visualized in PCA space. The Spearman's rank correlations between the descriptors were subjected to principal component analysis (PCA): ~74% of the accumulative variance is captured in the first three principal components shown here; therefore, descriptors located close to each other show stronger correlation based on the Spearman's rank-correlation coefficient. Overall, it can be observed that two clusters of descriptors emerged that show high correlation within their groups: first, the fingerprint-based descriptors ECFP4 and FCFP4, and, second, the pharmacophore-based descriptors TAT, TGT, TAD and TGD. In addition, MACCS keys and GpiDAPH3 descriptors did not show any significant correlation with other type of descriptors, nor did the descriptors BCUT, PMI, and prop2D.



Principal Components	% Explained Variance	% Accumulative Variance
1	49%	49%
2	15%	64%
3	11%	74%
4	9%	83%
5	7%	90%
6	4%	94%
7	3%	97%
8	2%	99%
9	1%	100%

Figure 9. Scree plot of the PCA of the 13-dimensional descriptor space (compared to the 37-dimensional descriptor space in an earlier study by Bender *et al.*[62]). The first three principal components capture 74% of the total variance, while five principal components are required to capture 90% of the total variance. In an earlier study by Bender *et al.*,[62] the first three components captured only 50% of the total variance, while 10 principal components were required to capture 80% of this measure. It cannot be definitely concluded whether more variance is captured in fewer dimensions, because of more similar behavior of descriptors or simply their lower number; however, it can be seen that a small number of dimensions already is sufficient to capture similarities and dissimilarities of the molecular descriptors when it comes to molecular diversity assessment, as employed in this study. When compared to the study by Bender *et al.*,[62] the increase in the percentage of accumulative explained variance for the first five principal components is very similar (41% here, compared to 38%), indicating that the first part of the curve is similar in both cases, albeit with a different percentage of variance, explained by the first principal component (49% here, compared to only 24% in the study by Bender *et al.*[62]).

It can be seen in **Figure 7** that the pharmacophore-based descriptors TAT, TAD, TGD, and TGT show a strong correlation with each other, with all Spearman's rank correlation coefficients observed being 0.74 (among TGT and TAT) or higher, indicating very similar behavior of those descriptors in diversity selection procedures. Accordingly, they were found to cluster together in **Figure 8**. The fifth pharmacophore-based descriptor utilized here, namely, GpiDAPH3, showed lower correlation with the previously mentioned descriptors, with Spearman's rank correlation coefficients observed with the descriptors TAD, TGD, and TGT of 0.46, 0.47, and 0.52, respectively, with the exception being TAT, which had a correlation coefficient with GpiDAPH3 of 0.65. This result is not surprising, considering that both of the latter descriptors capture pharmacophoric triangles (as

opposed to TAD and TGD, which use pharmacophore points), with GpiDAPH3 taking into account three-point pharmacophore fingerprints calculated from molecular graphs, and TAT taking into account atom-typed triangles calculated from the 3D conformation of a molecule. In addition, the pharmacophore-based GpiDAPH3 descriptor showed intermediate correlations between 0.54 and 0.67 with the fingerprint-based descriptors ECFP4 and FCFP4 and MACCS structural keys. The rest of the descriptors showed correlations of 0.51 or lower with the pharmacophore-based descriptors, indicating low similarity in behavior.

The fingerprint-based descriptors ECFP4 and FCFP4 showed a Spearman's rank correlation coefficient of 0.89 with each other, indicating stronger correlation with each other than the pharmacophore-derived descriptors, as is also clearly visible in **Figure 8**. This high correlation can be explained by the fact that both ECFP4 and FCFP4 are both derived from radial atom connectivity, and their similar behavior has been observed before in the context of similarity assessment.[62] In addition, MACCS structural keys have been shown to be correlated with ECFP4 and FCFP4 with a Spearman's rank correlation coefficient of 0.69 and 0.68, respectively, which can be explained by the local nature of both descriptor types. Rather surprisingly, the descriptors ECFP4 and GpiDAPH3 demonstrated a Spearman's rank correlation of 0.62, indicating that these two types of fingerprints demonstrate relatively similar behavior, which is more difficult to rationalize, given that the GpiDAPH3 descriptor takes into account graph-based pharmacophoric representations, as opposed to 2D structural features or atom type or counts. On the other hand, this is still lower than the correlation between TAT and GpiDAPH3 descriptors, so the relative ordering of descriptor pairs remains consistent with our initial expectations. The results discussed here indicate that molecular descriptors derived from atom topology or graph-based pharmacophoric representations tend to behave rather similarly overall, as shown in detail in **Figure 7** and **Figure 8**.

When now visiting descriptors of very different nature, we can see that overall PMI shows Spearman's rank correlations of 0.22 or lower with other type of descriptors, indicating very different behavior, whereas for ROCS, the Spearman's rank correlation coefficient with all other descriptors was also 0.55 or lower. ROCS and PMI also demonstrated significantly different behavior from each other, as measured by the Spearman's rank correlation coefficient, which was found to be only 0.22 (see **Figure 7** and **Figure 8**), even

though they both capture molecular shapes. This can be attributed to the fact that while PMI project molecular shapes (i.e., how similar molecules are to archetypal shapes such as spheres, disks, and rods) and are size-independent, ROCS considers not only molecular shape by comparing molecular shapes by overlapping Gaussian volumes, but also chemical similarity/atom types, thus being size-dependent. Similar differences exist for other descriptor types, where, e.g., 2D physicochemical descriptors, such as atom counts, molecular weight, and polar surface area, tend to increase as the size of the molecule increases and, hence, are often inherently size-dependent, while molecular fingerprints encode only the presence or absence of chemical substructures, and accordingly pay less attention to size.

As a nonstructural descriptor considered here, Bayes Affinity Fingerprints describe molecules by their predicted bioactivity spectra, and they showed low correlation with the rest of the descriptors utilized in this study: Spearman's rank correlations with all other descriptors did not exceed 0.53 in the case of FCFP4 and MACCS keys, and it was as low as 0.28 with GpiDAPH3 pharmacophores.

In order to further illustrate the extent to which different descriptors assess different aspects of chemical diversity, an example of comparison between two molecules in the Da library is given in **Figure 10**, where the ranks are shown for each descriptor employed (lower ranks mean higher similarity, whereas higher ranks mean lower similarity).

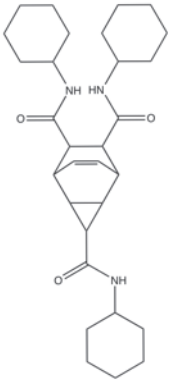
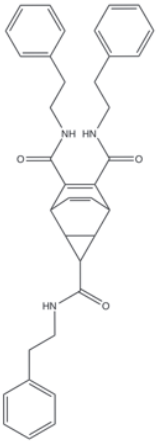
Query molecule	Target molecule	Descriptor	Ranking (%)
		2D properties	93
		BCUT	89
		Bayes affinity fingerprints	41
		ECFP4	26
		FCFP4	26
		MACCS	11
		GpiDAPH3	100
		PMI	26
		ROCS	100
		TAD	19
		TAT	11
		TGD	11
		TGT	11

Figure 10. Two molecules from the Da library that showed significant differences in ranking positions obtained from different descriptors used. These molecules were selected from the

Da library, which was shown to contain multiple molecules with antibacterial properties.[41,42] Pharmacophore-based and atom environments fingerprint descriptors showed very similar results, most likely because they encode only the presence or absence of chemical features, but do not take into account the size and/or molecular surface of the molecules, with the exception of GpiDAPH3. On the other hand, shape-based and 2D descriptors perceive the two molecules as significantly different, most likely because they take into consideration the entire molecular shape, hereby perceiving larger molecules as different from smaller ones (even if they consist of similar substructures). Hence, two chemical compounds subjectively perceived to be chemically similar by medicinal chemists could be considered as very similar, moderately similar, or even very different according to the descriptors used to assess chemical diversity.

It can be observed that, overall, there are significant differences among ranking positions obtained (ranks range from 3 to 27, given the library size of 27). MACCS keys and the pharmacophore-based descriptors TAT, TGT, TAD, and TGD showed very similar results (they all ranked the molecules with in the library at similar positions), and the fingerprint-based descriptors ECFP4 and FCFP4 also showed very similar results, compared among themselves (they both assigned a moderate similarity ranking).

However, the descriptors GpiDAPH3, ROCS, prop2D all perceived the molecules considered to be highly dissimilar. One possible explanation for the results observed above is that the descriptor MACCS was originally designed for indexing molecules and contains a small key set comprised of only 166 bits, which is the public subset (corresponding to predefined chemical functionalities), and it therefore appeared to be unable to distinguish well between relatively more similar molecules (much the same as the pharmacophore-based descriptors TAT, TGT, TAD, and TGD). However, the fingerprint-based descriptors ECFP4 and FCFP4 do take into account fingerprint features that are present in each molecule (independent of any predefined keyset), and they are hence able to also identify more subtle dissimilarities between those overall similar molecules. On the other hand, the descriptors prop2D and ROCS take into account substructural feature counts, molecular shape, and atom connectivity, and hence aggregate differences among the compounds, resulting in a perceived significance, with respect to physicochemical properties and shape of the structures. Hence, our findings illustrate that two molecules subjectively perceived to be chemically similar by visual inspection could be considered as highly similar, moderately similar, or even extremely different, according to the descriptors used to assess chemical diversity.

A comparison of descriptors on multiple datasets revealed that, although the Spearman's rank correlations between descriptors for larger datasets (such as PubChem, see **Figure 11**) resemble the correlations between descriptors for the compounds averaged across all libraries, analyses for particular datasets do not always show the same trend.

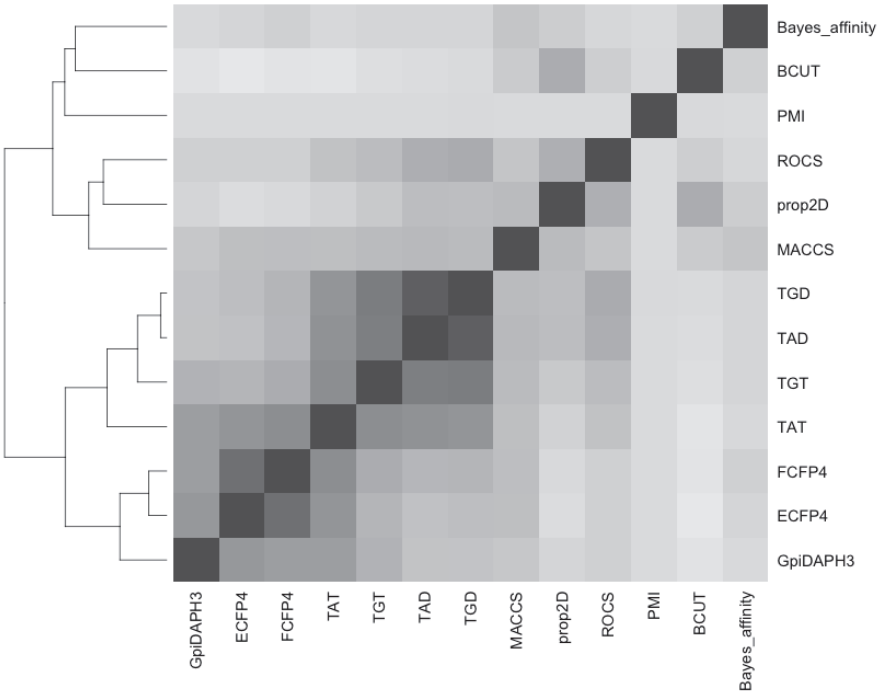


Figure 11. Spearman's rank correlation coefficient based on the overall averaged results over the PubChem library. Darker colors show higher correlation, whereas lighter colors show lower correlation among molecular descriptors. Overall, it can be seen that the Spearman's rank correlation pattern for the PubChem library is very similar to the overall Spearman's rank correlation pattern averaged over all libraries used in this study (recall **Figure 7**), with fingerprint-based and pharmacophore-based descriptors showing the most correlation among each other and within their respective group.

It can be seen that the descriptors correlate better for the DRS dataset (see **Figure 12**), where the average correlation for the DRS dataset across all descriptors was 0.53, whereas, for the PubChem dataset, this was only 0.31 (see **Figure 11**).

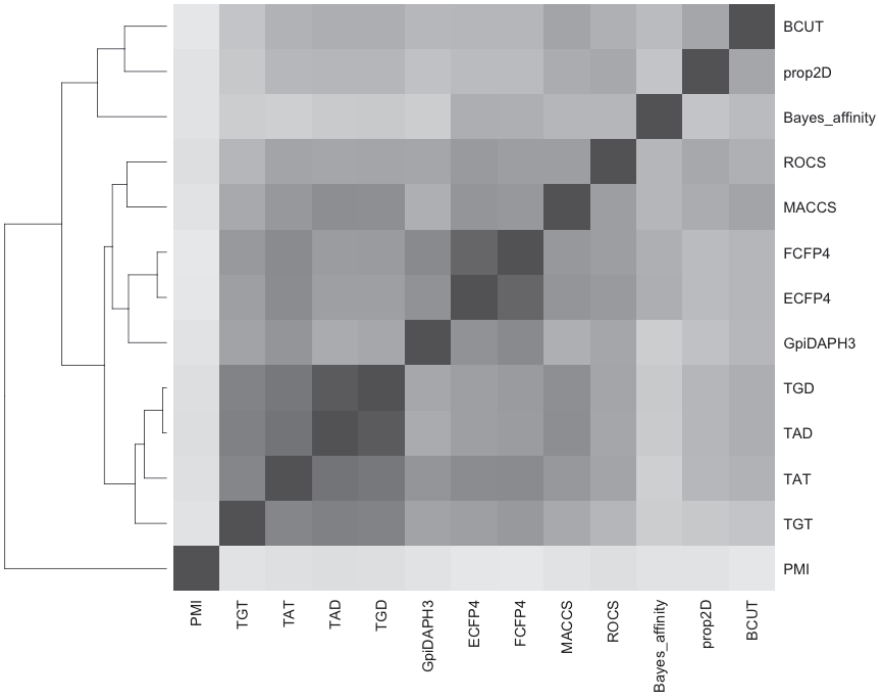


Figure 12. Spearman's rank correlation coefficient based on the overall averaged results over the DRS library. Darker colors show higher correlation, whereas lighter colors show lower correlation among descriptors. It can be observed that the descriptors correlate better for the DRS dataset than for the much larger PubChem dataset (see **Figure 11**): the average correlation for the DRS dataset across all descriptors was 0.53, whereas, for the PubChem dataset, this was only 0.31.

The Bayes Affinity Fingerprints show more similarity to fingerprint-based descriptors (which could be explained by the fact that this descriptor type was generated by a model trained on fingerprint-based descriptors) and shape-based descriptors, whereas PMI shows very low correlation with all other descriptors. Some descriptors show much higher correlations with other descriptors in the DRS dataset than either the PubChem or the overall dataset. For example, BCUT correlates very poorly with other descriptors (with an average correlation of 0.19) for the PubChem dataset, whereas for the DRS dataset, it shows a higher correlation of 0.49. Similarly, ROCS correlates poorly with other descriptors for the PubChem dataset (0.28), whereas for the DRS dataset, the average correlation was twice as high (0.57). These findings illustrate that the behavior of descriptors is highly dependent on the dataset analyzed and therefore, size and chemical composition of datasets should be taken into account when interpreting chemical diversity.

In order to correlate our findings with previous related studies, we compared our findings with the main results in the study previously reported by Bender *et al.*,[15] where PCA of the molecular descriptor space was performed, with respect to the similarity of molecules, with the objective being to understand which descriptors contain orthogonal information and which descriptors are correlated with each other. Many agreements and disagreements were observed regarding the correlation among the descriptors used between our PCA (**Figure 8**) and the PCA conducted by Bender *et al.*[15] First, fingerprint-based descriptors such as ECFP and FCFP cluster together in both PCA plots, as do the 3D pharmacophore-based descriptors TAT and TGT. However, the pharmacophore-based descriptor GpiDAPH3 is positioned away from other pharmacophore-based descriptors, but closer to the fingerprint-based descriptors in the current study, whereas, in the previous analysis by Bender *et al.*,[15] it has been located near other pharmacophore-based descriptors. In addition, MACCS keys (referred to as MDL in the study by Bender *et al.*[15]) are correlated with fingerprint-based descriptors in our study; however, this is not the case in the study by Bender *et al.*,[15] where MACCS keys are more correlated to pharmacophore-based descriptors instead. Differences observed in this study compared to the previous study reported can be attributed to different objectives, because, in this study, the objective was to evaluate the correlation of molecular descriptors based on rankings obtained from calculated distances among compounds present in chemical libraries for diversity assessment, instead comparing the performance of molecular descriptors in retrieving active compounds for virtual screening assessment.

In order to assess not only similarities and dissimilarities in the behavior of different descriptors when applied to diversity selection but also their *performance, with respect to a relevant measure*, eight out the 13 descriptors were used for diversity selection and their performance was assessed by the coverage of protein targets in bioactivity space. (Some methods could not be used due to computational demands, given that the full compound similarity matrix needed to be computed for diversity selection using the methods employed here.) The performance of each method was assessed based on the number of activity classes covered by sampling a diverse subset of 4% (100 compounds) from an initial set of 2587 randomly selected compounds covering the 25 largest ChEMBL activity classes of human protein targets. These 25 activity classes contain 103 data points on average and vary in size

from 12 data points to 190 data points, indicating that this is an unbalanced dataset.

Results of this analysis are shown in **Table 4**. Bayes Affinity Fingerprints, considering only predicted protein targets with a Bayes score above 30 ("Cutoff 30"), sampled an average of 92% of bioactivity classes, hence outperforming all other descriptors marginally. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes, respectively, were represented in the selected compounds. Random sampling retrieved ~84% of activity classes (averaged over three attempts), showing similar performance on this dataset to the descriptors GpiDAPH3, TGT, and BCUT, while outperforming molecular descriptors such as prop2D (80%), MACCS (80%), and PMI (75%). Despite the seemingly high performance of random sampling, it should be noted that random selection would only yield the best results (compared to any other method based on molecular descriptors) in the case where all classes are of equal size by picking up the most diverse subsets in bioactivity space. This could lead one to the misconception that random selection is a better option than the currently available methods used for diversity selection. However, in more realistic situations, such as here, some biological targets are more promiscuous than others, and hence they can accommodate a more diverse set of compounds in their binding sites than others (therefore leading to "larger activity classes" against those more promiscuous proteins). In such cases, random selection would struggle to pick compounds from small classes and, thus, does not seem to be the most suitable approach to be applied for diversity selection. Diversity selection methods based on molecular descriptors appear to be less successful in retrieving bioactive compounds against a broad range of protein targets, since no prior knowledge of which bits in the fingerprint matter (and lead to bioactivity differences) is considered by these methods. Instead, Bayes Affinity Fingerprints, which are trained on active compounds covering a large part of chemogenomic space, subselect compounds by taking into account known bioactive chemistry (and modifications leading to bioactivity changes), and thus appear to be able to achieve better protein target coverage. Our results are consistent with our previously reported study by Nguyen et al.,[24] where diversity selection based on bioactivity spectra fingerprints outperformed commonly employed circular fingerprint-based methods by up to 10%, when sampling bioactive compounds.

Finally, we attempted to assess whether PMI plots, which have frequently been used to assess the diversity of, e.g., DOS libraries, can be used to this end, when also paying attention to bioactivity coverage. This analysis is presented in **Figure 13**, as a PMI ratios plot (nPR1 and nPR2) for 6 out of 50 ChEMBL activity classes.

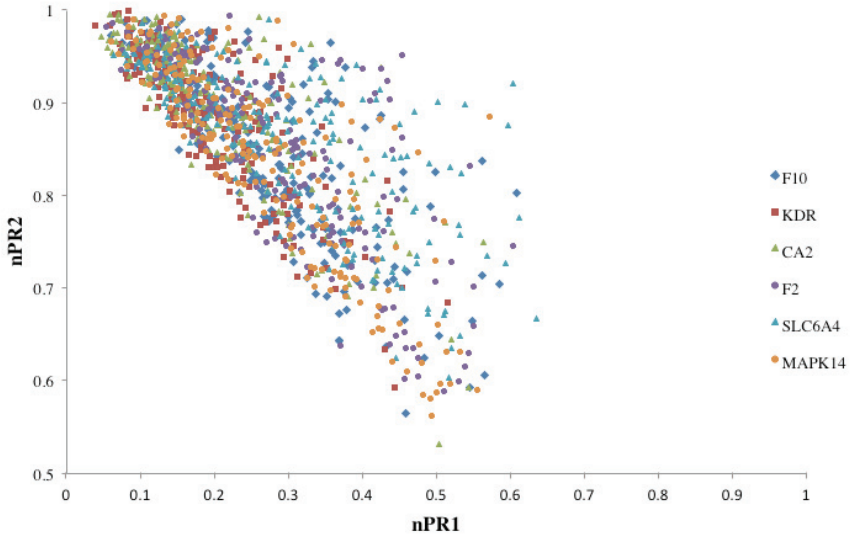


Figure 13. Normalized PMI ratios (nPR1 and nPR2) plot of 6 out of 50 ChEMBL activity classes utilized; namely, the coagulation factor X (F10), vascular endothelial growth factor receptor 2 (KDR), carbonic anhydrase 2 (CA2), prothrombin (F2), sodium-dependent serotonin transporter (SLC6A4), and mitogen-activated protein kinase 14 (MAPK14). It can be seen that molecular shape diversity of chemical libraries, measured by PMI, does not correlate with or indicate diversity of coverage in bioactivity space, as compounds binding to different protein families occupy similar space, and, in return, compounds from different areas of PMI space are bioactive against the same protein. Hence, the authors would recommend that, although PMI analyses give an insight into the shape properties of a chemical library, they might not be the most suitable tool to assess compound diversity in bioactivity space.

It can be seen that molecular shape diversity of chemical libraries, as measured by PMI, does not correlate or indicate a diversity of coverage in bioactivity space, because compounds binding to different protein families occupy similar space, and compounds from different areas of PMI space are bioactive against the same protein. However, one could argue that the targets from ChEMBL represent a biased part of biological space (the space where it has been easy to identify bioactive molecules with traditional medicinal chemistry efforts). Therefore, a similar analysis on 1995 compounds with

annotated K_i, IC_{50}, EC_{50}, or K_d values of 10 μM or better from the TIMBAL database was performed,[43] which involves a database of compounds with protein−protein inhibitory properties. It also can be seen that, in this case, molecular shape diversity does not correlate or indicate diversity of coverage in bioactivity space, because the compounds have a similar distribution to that in **Figure 13**. Hence, in the authors' opinion, PMI analyses, while giving insight into the shape properties of a chemical library, should rather not be used to assess diversity in bioactivity space. One might find this assumption tempting to make; however, **Figure 13** illustrates the rather small correlation between diversity in PMI space and diversity in bioactivity space; moreover, from this illustration, it is apparent that diversity in one space has little predictive value for diversity in the other. In the opinion of the authors, diversity in bioactivity space – even if only computationally established – would be a practically more relevant measure for assessing the biologically relevant diversity of small molecule libraries.

Conclusions

This study aimed at assessing both the correlations between molecular descriptors in the context of diversity assessment, as well as their performance, with respect to covering bioactivity space. It was found that descriptors derived from atom topology (i.e., pharmacophore-based descriptors, such as TAT, TAD, TGD, and TGT) and fingerprint-based descriptors (such as ECFP4 and FCFP4) generally showed strong correlation within each group (all with Spearman's rank correlations of 0.74 or higher) and between both groups (all with Spearman's rank correlations of 0.59 or higher). On the other hand, shape-based descriptors such as rapid overlay of chemical structures (ROCS) and principal moments of inertia (PMI) demonstrated behaviors that were significantly different from each other, as measured by the Spearman's rank correlation coefficient, which was found to be only 0.22. Moreover, it was observed that descriptors correlate differently, depending on the dataset used. For example, the average correlation of descriptors for the DRS dataset, encompassing 28 dissimilar compounds (with multiple scaffolds being present in the dataset), is 0.53, whereas for the much larger and more diverse (on an absolute scale) PubChem dataset, the average correlation reached only 0.31. Hence, overall our results indicate that molecular descriptors differ in the way they assess chemical diversity, depending on the diversity and size of the datasets used, and selecting the

appropriate descriptor is a nontrivial task, which must take into account both of these aspects. Moreover, the shape-based descriptor PMI showed no correlation with any other type of descriptor utilized in this study, demonstrating very different behavior from all other descriptors employed here.

Given that diversity in bioactivity space is often the primary objective of compound diversity selection procedures, we furthermore benchmarked the descriptors employed here, with respect to covering 25 bioactivity classes upon selecting 4% of 2587 compounds from a subset of ChEMBL. Here, it was found that the Bayes Affinity Fingerprints showed the best performance by covering 92% of bioactivity classes, hence outperforming all other descriptors marginally. Using the descriptors ECFP4, GpiDAPH3, TGT, and random sampling, 91%, 84%, 84%, and 84% of the activity classes, respectively, were represented in the selected compounds, followed by BCUT, prop2D, MACCS, and PMI (in order of decreasing performance). Random sampling retrieved ~84% of activity classes (averaged over three attempts), showing similar performance of this dataset to the descriptors GpiDAPH3, TGT, and BCUT, while outperforming molecular descriptors such as prop2D (80%), MACCS (80%), and PMI (75%). In addition, we were able to show that there is no visible correlation between compound diversity in PMI space and in bioactivity space, despite frequent utilization of PMI plots to this end.

Overall, this study assessed which descriptors are able to select compounds with high coverage of bioactivity space, and which can hence be used for diverse compound selection for biological screening. It also gives guidelines as to which descriptors behave rather collinear and which descriptors behave more orthogonal in diversity selection tasks. We propose that a combination of complementary descriptors such as Bayes Affinity Fingerprints, PMI, and prop2D be used to computationally select a diverse set of compounds for screening purposes. Such a computational "filter" might also add value to current endeavors of assembling screening libraries against diverse targets, such as the European Lead Factory,[63] as well as library enhancement initiatives that are taking place continuously in pharmaceutical companies.

References

(1) Bohacek, R. S., McMartin, C., Guida, W. C. (1996) The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev. 16*, 3–50.

(2) Jorgensen, W. L. (2004) The many roles of computation in drug discovery. *Science 303*, 1813–1818.

(3) Huggins, D. J., Venkitaraman, A. R., Spring, D. R. (2011) Rational methods for the selection of diverse screening compounds. *ACS Chem. Biol. 6*, 208−217.
(4) Dobson, C. M. (2004) Chemical space and biology. *Nature 432*, 824−828.
(5) Lipinski, C., Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature 432*, 855−861.
(6) Maggiora, G. M., Johnson, M. A. (1990) *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York.
(7) Bender, A., Glen, R. C. (2004) Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem. 2*, 3204−3218.
(8) Perez, J. J. (2005) Managing molecular diversity. *Chem. Soc. Rev. 34*, 143−152.
(9) Petrone, P. M., Wassermann, A. M., Lounkine, E., Kutchukian, P., Simms, B., Jenkins, J., Selzer, P., Glick, M. (2013) Biodiversity of small molecules – A new perspective in screening set selection. *Drug Discov. Today 18*, 674−680.
(10) Willett, P. (1999) Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol. 6*, 447−457.
(11) Roth, H. J. (2005) There is no such thing as "diversity"! *Curr. Opin. Chem. Biol. 9*, 293−295.
(12) Kutchukian, P. S., Vasilyeva, N. Y., Xu, J., Lindvall, M. K., Dillon, M. P., Glick, M., Coley, J. D., Brooijmans, N. (2012) Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery. *PloS One 7*, e48476.
(13) Lajiness, M. S., Maggiora, G. M., Shanmugasundaram, V. (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem. 47*, 4891−4896.
(14) Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., Weinberger, L. E. (1996) Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem. 39*, 3049−3059.
(15) Bender, A. (2010) How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discovery 5*, 1141−1151.
(16) Duan, J., Dixon, S. L., Lowrie, J. F., Sherman, W. (2010) Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell. 29*, 157−170.
(17) Fergus, S., Bender, A., Spring, D. R. (2005) Assessment of structural diversity in combinatorial synthesis. *Curr. Opin. Chem. Biol. 9*, 304−309.
(18) Akella, L. B., DeCaprio, D. (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol. 14*, 325−330.
(19) Clemons, P. A., Wilson, J. A., Dancik, V., Muller, S., Carrinski, H. A., Wagner, B. K., Koehler, A. N., Schreiber, S. L. (2011) Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl. Acad. Sci. U.S.A. 108*, 6817−6822.
(20) Naylor, E., Arredouani, A., Vasudevan, S. R., Lewis, A. M., Parkesh, R., Mizote, A., Rosen, D., Thomas, J. M., Izumi, M., Ganesan, A., Galione, A., Churchill, G. C. (2009) Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol. 5*, 220−226.
(21) McGregor, M. J., Muskal, S. M. (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci. 39*, 569−574.
(22) Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today 11*, 1046−1053.
(23) Grant, J. A., Gallardo, M. A., Pickup, B. T. (1996) A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem. 17*, 1653−1666.
(24) Nguyen, H. P., Koutsoukas, A., Mohd Fauzi, F., Drakakis, G., Maciejewski, M., Glen, R. C., Bender, A. (2013) Diversity selection of compounds based on "Protein Affinity

Fingerprints" improves sampling of bioactive chemical space. *Chem. Biol. Drug. Des. 82*, 252−266.
(25) Gillet, V. J. (2011) Diversity selection algorithms. *Wiley Interdiscip. Rev.: Comput. Mol. Sci. 1*, 580−589.
(26) Lajiness, M. S. (1990) Molecular similarity-based methods for selecting compounds for screening. In *Computational Chemical Graph Theory*; Nova Science Publishers, Inc.: Commack, NY, USA; pp 299−316.
(27) Lipkowitz, K. B., Boyd, D. B. (2003) Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*, Vol. 18; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2003.
(28) Lewis, R. A., Mason, J. S., McLay, I. M. (1997) Similarity measures for rational set selection and analysis of combinatorial libraries: The Diverse Property-Derived (DPD) approach. *J. Chem. Inf. Comput. Sci. 37*, 599−614.
(29) Hassan, M., Bielawski, J. P., Hempel, J. C., Waldman, M. (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity 2*, 64−74.
(30) Waldman, M., Li, H., Hassan, M. (2000) Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graphics Modell. 18*, 412−426, 533−536.
(31) Bender, A., Jenkins, J. L., Glick, M., Deng, Z., Nettles, J. H., Davies, J. W. (2006) "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. 46*, 2445−2456.
(32) Galloway, W. R., Isidro-Llobet, A., Spring, D. R. (2010) Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun. 1*, 80.
(33) Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C.Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., NeveuV., Greiner, R., Scalbert, A. (2013) HMDB 3.0 – The Human Metabolome Database in 2013. *Nucleic Acids Res. 41*, D801−D807 (Database Issue).
(34) Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M. (2008) DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res. 36*, D901−D906 (Database Issue).
(35) Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Bryant, S. H. (2009) PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res. 37*, W623−W633 (Web Server Issue).
(36) Gaulton, A., Bellis, L. J., BentoA. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J. P. (2012) ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res. 40*, D1100−D1107 (Database Issue).
(37) Spandl, R. J., Bender, A., Spring, D. R. (2008) Diversity-oriented synthesis; a spectrum of approaches and results. *Org. Biomol. Chem. 6*, 1149−1158.
(38) Beckmann, H. S. G., Nie, F., Hagerman, C. E., Johansson, H., Tan, Y. S., Wilcke, D., Spring, D. R. (2013) A New Strategy for the Diversity-Oriented Synthesis of Macrocyclic Scaffolds using Multi-Dimensional Coupling. *Nat. Chem. 5*, 861−867.
(39) Mullard, A. (2012) Protein-protein interaction inhibitors get into the groove. *Nat. Rev. Drug Discovery 11*, 173−175.
(40) Wolfson, W. (2012) Grabbing for the ring: macrocycles tweak the conventions of drug making. *Chem. Biol. 19*, 1356−1357.
(41) Wyatt, E. E., Fergus, S., Galloway, W. R., Bender, A., Fox, D. J., Plowright, A. T., Jessiman, A. S., Welch, M., Spring, D. R. (2006) Skeletal diversity construction via a branching synthetic strategy. *Chem. Commun. (Cambridge, U. K.) 31*, 3296−3298.

(42) Wyatt, E. E., Galloway, W. R., Thomas, G. L., Welch, M., Loiseleur, O., Plowright, A. T., Spring, D. R. (2008) Identification of an anti-MRSA dihydrofolate reductase inhibitor from a diversity-oriented synthesis. *Chem. Commun. (Cambridge, U. K.) 40*, 4962−4964.
(43) Higueruelo, A. P., Schreyer, A., Bickerton, G. R., Pitt, W. R., Groom, C. R., Blundell, T. L. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des. 74*, 457−467.
(44) *ChemAxon Standardizer*, version 5.12; ChemAxon, Ltd: Budapest, Hungary, 2012.
(45) *Molecular Operating Enviroment (MOE)*, version 2011.10; Chemical Computing Group, Inc: Montreal, Canada, 2012.
(46) Durant, J. L., Leland, B. A., Henry, D. R., Nourse, J. G. (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci. 42*, 1273−1280.
(47) Rogers, D., Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model. 50*, 742−754.
(48) Morgan, H. L. (1965) The Generation of a Unique Machine Description for Chemical Structures – A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc. 5*, 107−113.
(49) Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A. (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem. 2*, 3256−3266.
(50) Bender, A., Mussa, H. Y., Glen, R. C., Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci. 44*, 1708−1718.
(51) Williams, C. (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Diversity 10*, 311−332.
(52) Rush, T. S., 3rd, Grant, J. A., Mosyak, L., Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem. 48*, 1489−1495.
(53) *OEChem vROCS*, version 3.1.2; OpenEye Scientific Software: Santa Fe, NM, USA, 2011.
(54) Sauer, W. H., Schwarz, M. K. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci. 43*, 987−1003.
(55) Pearlman, R. S., Smith, K. M. (1999) Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci. 39*, 28−35.
(56) Koutsoukas, K., Lowe, R., KalantarMotamedi, Y., Mussa, H. Y., Klaffke, W., Mitchell, J. B. O., Glen, R. C., 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.
(57) Deza, M. M. *Encyclopedia of Distances*, Second Edition; Springer: New York, 2012.
(58) Hamming, R. W. (1950) Error detecting and error correcting codes. *Bell System Tech. J. 29*, 147−160.
(59) Myers, J. L., Well, A., Lorch, R. F. (2010) *Research Design and Statistical Analysis*, Third Edition; Routledge: New York, 2010; 809 pp.
(60) R Core Team. R: *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010.
(61) *Dotmatics Vortex*, version 2013.03.20719, Dotmatics: The Old Monastery, Windhill, Bishops Stortford, Herts, U.K., 2013.
(62) Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C., Glick, M., Davies, J. W. (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model. 49*, 108−119.
(63) Mullard, A. (2013) European lead factory opens for business. *Nat. Rev. Drug Discovery 12*, 173−175.

Chapter four

Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data

Abstract

With increased automation and larger compound collections, the development of high-throughput screening (HTS) started replacing previous approaches in drug discovery from around the 1980s onward. However, even today it is not always appropriate, or even feasible, to screen large collections of compounds in a particular assay. Here, we present an efficient method for iterative screening of small subsets of compound libraries. With this method, the retrieval of active compounds is optimized using their structural information and biological activity fingerprints. We validated this approach retrospectively on 34 Novartis in-house HTS assays covering a wide range of assay biology, including cell proliferation, antibacterial activity, gene expression, and phosphorylation. This method was employed to retrieve subsets of compounds for screening, where selected hits from any given round of screening were used as starting points to select chemically and biologically similar compounds for the next iteration. By only screening ~1% of the full screening collection (~15000 compounds), the method consistently retrieves diverse compounds belonging to the top 0.5% of the most active compounds for the HTS campaign. For most of the assays, over half of the compounds selected by the method were found to be among the 5% most active compounds of the corresponding full-deck HTS. In addition, the stringency of the iterative method can be modified depending on the number of compounds one can afford to screen, making it a flexible tool to discover active compounds efficiently.

Introduction

Early drug discovery traditionally has been the result of a close collaboration between chemists, pharmacologists, and clinical scientists, where knowledge from pharmacology and (medicinal) chemistry was combined to design potentially active molecules for testing.[1,2] From around the 1980s onward, rapid improvements in automation and combinatorial chemistry led to the development and increasing acceptance of high-throughput screening (HTS), which allows rapid screening of large collections of compounds using robotics and automated data processing. This enabled HTS to be used to study relationships between compounds and putative biological targets on a very large scale, so that libraries of 1−2 million compounds are routinely screened in big pharmaceutical companies, several times per year.[2,3] Conceptually, HTS aims to screen large numbers of molecules in a brute-force

approach to identify hits, and the most promising chemical entities are then selected as starting points for further investigation. It is hoped that screening large numbers of molecules increases the chance of finding promising chemical entities. However, the previous often iterative cycles of design–screen–refine in small interdisciplinary project teams were somewhat lost.

Over the past few decades, HTS has hence become increasingly popular and has been augmented in capacity from being able to screen tens of thousands of compounds a day to over 100000 compounds a day and has become – in addition to many other techniques – of crucial importance for early drug discovery.[4–6] However, HTS also has some significant drawbacks. Cell-free HTS campaigns, such as biochemical target-based assays, are not adequately predictive of compounds' ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, which are important pharmacokinetic parameters for drug development.[7] For cell-based phenotypic HTS assays, which can be more predictive of certain ADMET properties such as bioavailability and cytotoxicity, target deconvolution is an important challenge.[8] Additionally, HTS campaigns sometimes cannot be performed at scale for complex biological systems that cannot be mass-produced (e.g., organoids).[9] Finally, and of most relevance for the current study, HTS remains a resource-intensive endeavor with a large fraction of the compounds screened being inactive or uninteresting. The latter renders the identification of smaller screening sets which lead to a significant fraction of active chemical matter detected very relevant.[4]

The mentioned drawbacks prompted efforts to optimize various aspects of HTS campaigns, such as compound library design (for example, based on chemical diversity, where libraries are chosen on the basis of chemical knowledge),[10–14] post-HTS data analysis for triaging active compounds (in order to select subsets for further validation),[15,16] and selecting novel compounds similar to active compounds detected in the assay for further investigation.[17–21] Given the recent perceived ineffectiveness of target-based HTS,[22] a shift to phenotypic HTS has occurred,[8] hence increasing the need for target identification methods. In this regard, a high-throughput screening fingerprint (HTS-FP) capturing past performance of compounds across a number of screens was developed by Petrone *et al.* at Novartis,[23] which allows the comparison of compounds according to their bioactivity across a range of HTS assays. This approach was used for both similarity searching and various

machine learning methods for target identification of hits from phenotypic screens. Later, a public version of the same fingerprint was developed and analyzed by Dančík *et al.*,[24] who also reported its usefulness in the elucidation of the compound mode of action. However, despite these computational advances in postscreen analysis, HTS campaigns remain an expensive endeavor.

In this study, we aim to address efficient ways of screening subsets of compound libraries, instead of screening entire compound libraries, while at the same time optimizing the retrieval of active compounds. We developed and retrospectively validated an iterative screening method on Novartis in-house HTS data, in which selected hits from any given round of screening were used as starting points to select chemically and biologically similar compounds for the next iteration. This approach was developed with the explicit aim to select much smaller subsets of compounds with enriched activity, by harnessing the bioactivity information on compounds in the previous iteration. While briefly mentioned by Mayr and Bojanic as an idea,[5] and used on a small scale by Keenan *et al.* for the design of plasmodial kinase inhibitors,[25] the concept of iterative screening has not been explored systematically in the published literature. A related concept has been previously described by Schneider *et al.* in the context of iterative virtual synthesis and testing of individual molecules, where molecules are designed automatically using evolutionary algorithms and particle swarm optimization.[26] However, our approach differs considerably, because we iteratively generate sets of molecules instead of individual molecules, hence investigating the concept on a much larger scale.

Methods

HTS Data. Novartis proprietary HTS assays comprising at least 1300000 compounds with an inhibitory assay readout were used, resulting in a total of 34 assays, of which 11 were cell-based assays and 23 were cell-free (biochemical) assays. These assays covered a wide number of biological events, including cell proliferation, antibacterial activity, gene expression, and phosphorylation.

Starting Set for Initial Screening Round. We used a starting set of well-studied and manually curated compounds, many with tested clinical relevance, known to cover a large amount of druggable bioactivity space and of which the mechanism of action (MoA) is known. This set (the MoABox)

comprised 2757 compounds and is used as a starting point for many phenotypic screening projects at Novartis due to the high-quality annotations of each compound. The physicochemical properties and the chemical and biological diversity of the MoABox were calculated using RDKit[27]. The design of the MoABox inherently entails that most compounds have properties favorable for cell-based screening. Owing to operational turnover of the compound archive, not every full-deck HTS contains every compound of the MoABox. Therefore, the starting set for each specific assay was the MoABox compounds present in it at the time it was performed. The smallest starting set comprised 2050 compounds, whereas the largest comprised 2692 compounds.

In order to determine the importance of the starting set for good performance, we repeated our analysis with 10 randomly chosen starting sets and the results were compared to those obtained with the MoABox as a starting set. These sets were obtained by repeatedly selecting a random subset from the entire screening deck of equal size to that of the MoABox present in the corresponding assay, minus any MoABox compounds that might have been coincidentally selected.

Iterative Screening Algorithm (ISA). For any given set of compounds, we are able to look up its activities in a past assay with ~1.3 M compounds. This *in silico* screening allows not only a relative ranking (according to activities within the subset) but also an absolute ranking (according to the 1.3 M compounds). Our aim was to iteratively optimize the absolute ranking of subsets of compounds, thereby efficiently selecting highly active compounds and steering the screening process toward success with much smaller compound sets. Therefore, the method developed in this study consists of three iterative procedures (see **Figure 14**): (1) ranking of compounds based on retrospective activity data, (2) selection/triaging of hits, and (3) expanding from hits to close analogs based on chemical and biological similarity metrics.

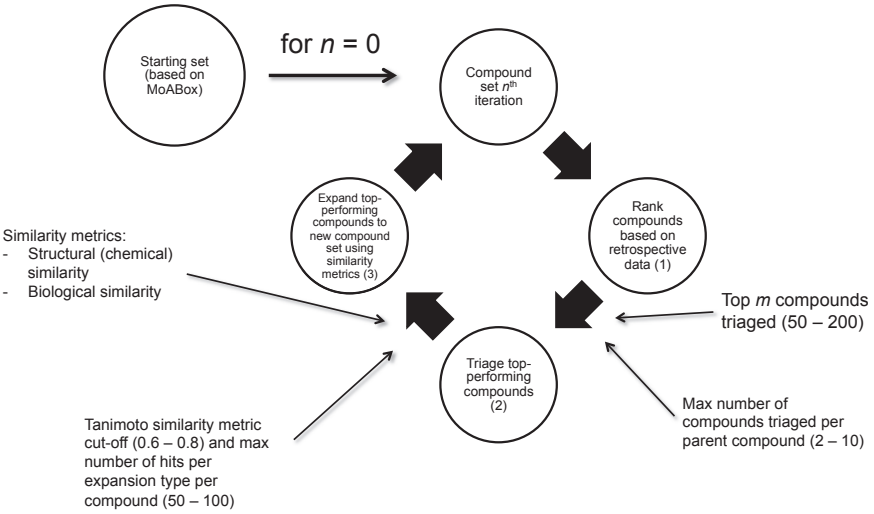


Figure 14. Iterative screening algorithm (ISA) overview. The ISA developed in this study consisted of three iteration steps: (1) ranking of compounds based on retrospective data, (2) triaging of (i.e., selecting) top-performing compounds, and (3) expanding from top-performing compounds to close analogs based on chemical and biological similarity metrics. The starting set comprises the MoABox compounds present in the HTS assay. The ISA allows for adjustment of parameters at the triaging stage (the number of compounds carried forward, and the number of compounds originating from the same parent compound to limit large numbers of closely related analogs). At the expansion stage, the parameters used (chemical and/or biological similarity) can be adjusted, as well as the corresponding similarity cutoff and maximum number of expansions per compound.

Since this study is a retrospective analysis on HTS data, the ranks of the compounds selected correspond to the ranks of the same compounds had they been screened in a full-deck screen. Our method is fundamentally different from a basic similarity search using active probes, because we perform a similarity search iteratively based on active compound information at every round of screening, rather than only once. Circular fingerprints[28] (SciTegic ECFP4-like) were used as features for determining chemical similarity, and HTS fingerprints (HTS-FP)[23] were used as features for determining biological similarity.

Metrics Used for Performance Assessment. We used two criteria for evaluating compound sets at each iteration: (1) the rank distribution based on compound activity and (2) the cumulative coverage of Murcko scaffolds[29] found in the top 0.5% of compounds ranked by activity. In conjunction, these criteria assess the retrieval of not only active but also structurally diverse sets of compounds. A median rank cutoff of 65000 is sometimes used to assess

performance; this corresponds to 5% of a total screening collection of 1.3 million compounds.

Systematic Exploration of Parameters. The number of compounds triaged per iteration as well as the number and types of expansions affect the size and diversity of the compound sets selected. First, the number of top-performing compounds triaged can be varied. Second, expansions can be adjusted (chemical and/or biological similarity), as well as the corresponding Tanimoto[30] similarity cutoff and maximum number of expansions per compound. Moreover, the maximum number of compounds originating from the same parent compound can be adjusted in order to limit the number of closely related analogs. We systematically explored the influence of these parameters in a number of *in silico* experiments (see **Table 5**), where the influence of each parameter was analyzed individually.

Table 5. Summary of parameters explored over nine *in silico* experiments. Here, experiment 1 was considered as the reference experiment, which was chosen on basis of a trade-off between number of compounds screened over 10 iterations (approximately 1% of screen size) and performance. All other experiments varied one parameter, therefore allowing an assessment of its influence with respect to the reference experiment. For example, a comparison of experiment 3 with experiment 1 shows the effect of doubling (from 100 to 200) the number of compounds triaged per iteration.

Experiment number	**Iteration count**	**Triaged number of compounds**	**Maximum number of expansions (structure-based)**
1	10	100	50
2	10	50	50
3	10	200	50
4	10	100	100
5	10	100	50
6	10	100	50
7	10	100	50
8	10	100	50
9	10	100	50
Tanimoto cutoff (structure-based)	**Maximum number of expansions (HTS-FP-based)**	**Tanimoto cutoff (HTS-FP-based)**	**Maximum number of compounds triaged per parent compound**
0.6	50	0.6	5
0.6	50	0.6	5
0.6	50	0.6	5
0.6	50	0.6	5
0.8	50	0.6	5
0.6	100	0.6	5
0.6	50	0.8	5

0.6	50	0.6	2
0.6	50	0.6	10

Experiment 1 was considered as a realistic reference experiment that balances performance and the number of compounds screened over 10 iterations (~1% of entire collection, ~15000 compounds). All other experiments varied one parameter, therefore allowing an assessment of its influence with respect to the reference experiment. For example, a comparison of experiment 3 with experiment 1 shows the effect of doubling the number of compounds triaged per iteration from 100 to 200.

Data Analysis. The workflow comprised Python and Perl scripts for data analysis and the Indigo toolkit[31] and RDKit[27] for cheminformatics calculations. Spotfire[32] was used for data exploration, and R[33] and Cytoscape[34] were used for the visualization of results.

Results and discussion

Here, we present in detail the results belonging to the reference experiment (experiment 1 in **Table 5**), followed by a comparison to other experiments. Experiments 4, 6, and 7 showed the same results as the reference experiment and are therefore not discussed separately; these experiments highlight, however, that more than 50 expansions or a more stringent HTS-FP similarity cutoff do not change the results.

Iterative Screening Is Highly Effective Across Assay Types. The median rank of the compounds selected was 36101 (excluding the starting set) across all assay types, which corresponds to the top ~2.8% of a collection of ~1.3 M compounds. In other words, half the compounds selected across all iterations (except for the starting set) are found among the top 2.8% of the corresponding 1.3 M compound screen, indicating a clear enrichment in activity of the compounds selected. Of note, the performance is consistent for a large number of different assay types (median rank below 65000, see **Figure 15**).

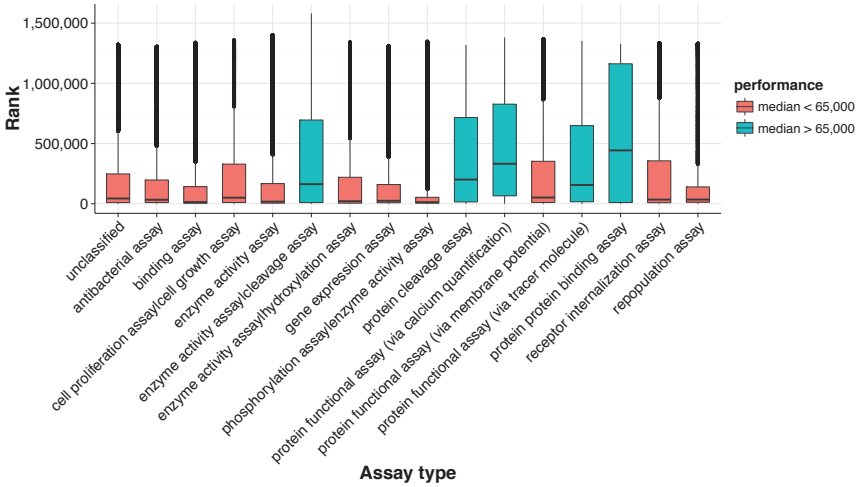


Figure 15. Ranks of compounds from iterations for all assay types. Boxplots of ranks for all compounds selected by the iterative screening algorithm (ISA) for iterations 1–10 (excluding the starting set) are represented for each assay type. The performance for enzyme activity/cleavage assay, protein cleavage assay, protein functional assay, and protein–protein binding assay is much worse (median rank of 200000 on average) compared to other assays, with also a broader rank distribution. Red, median rank below 65000; blue, median rank above 65000. The first 65000 compounds correspond to the top ~5% of 1.3 M.

However, for the types enzyme activity/ cleavage assay, protein cleavage assay, protein functional assay, and protein–protein binding assay, the performance was reduced, as evidenced by a median rank greater than 65000 combined with a higher standard deviation.

Interestingly, performance is better for the cell-free assays than the cell-based assays (rank distributions for both assay formats is shown in **Figure 16**).

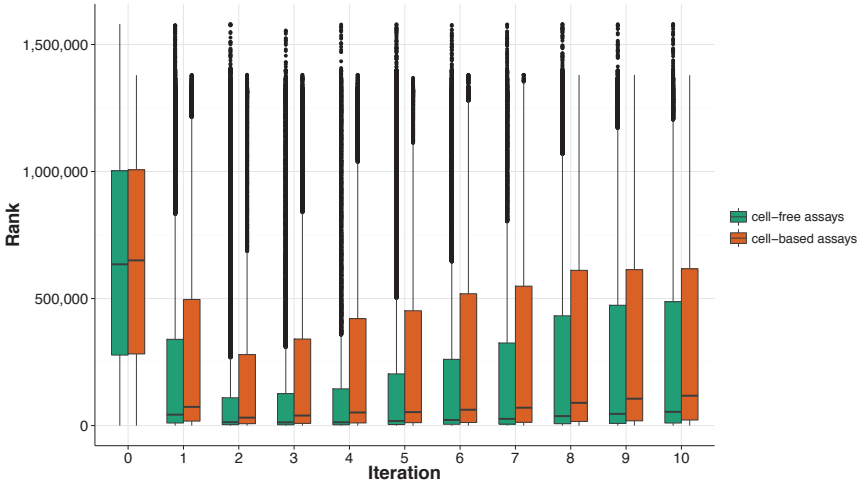


Figure 16. Ranks of iteratively selected compounds for cell-free and cell-based assays. Green, cell-free assays; orange, cell-based assays. There is a consistent difference in median rank (and interquartile range, extension of boxplot) across iterations 1 to 10 between cell-free and cell-based assays. This indicates the relative difficulty in selecting compounds that are able to satisfy cell-based screening requirements (e.g., cell permeability). Median ranks are significantly different (paired *t*–test, *p*-value < 10^{-5}), as are the rank distributions for each iteration (Kolmogorov–Smirnov test, *p*-value < 10^{-5}).

In order to investigate whether this difference was statistically significant, a paired *t*–test was performed for the median ranks across iterations 1 to 10. In addition, a Kolmogorov–Smirnov test was performed for every iteration on compound ranks of different assay formats. All *p*-values were smaller than 10^{-5}, hence indicating a statistically significant difference in distribution of rank between cell-free and cell-based assays. This difference is likely due to the fact that in order for compounds to have an effect in cell-based assays, they have to be able to cross the cell membrane to reach the target of interest (in cases when this target is not membrane-bound). Hence, these compounds must have suitable physicochemical properties (such as permeability), in order to be effective. Since our method on purpose did not distinguish between cell-free and cell-based assays, these results are in line with expectations; however, specific compound criteria for cell-based assays (e.g., incorporation of logP values, past performance in cell-based assays) are likely to diminish this observed gap in performance between the two assay formats in the future. Our starting set, the MoABox (see Methods), is geared toward hypothesis-generating cell-based phenotypic screening; as a result, this set of

compounds performs equally well on cell-based and cell-free assays (**Figure 16**, iteration 0).

Next, median compound ranks were evaluated per assay type (**Figure 17**).

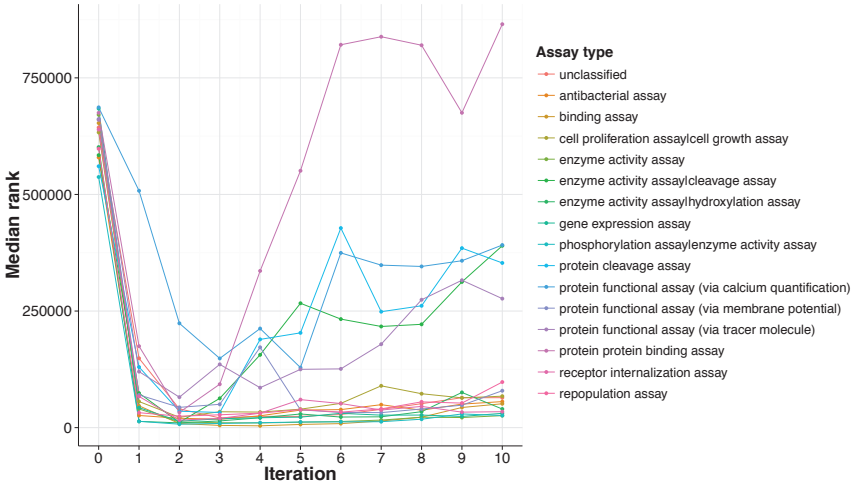


Figure 17. Median rank per iteration across assay types. The median rank of the compounds of the selected subset at each iteration is plotted versus iteration. The iterative screening algorithm (ISA) performs consistently well for most assay types, but there are a number of assays for which the median rank of compounds selected swiftly deteriorates after around iteration 3. These assays are for protein–protein binding, protein cleavage, protein function, and enzyme activity/cleavage and are the same ones shown to have an overall median rank greater than 65000 (**Figure 15**).

The iterative method performs consistently well for the majority of assay types (median ranks are smaller than 100000 for iterations 1–10 for 11 out of the 16 assay types), but there are a number of outlier assay types for which the median rank of compounds selected swiftly deteriorates after around iteration 3. These assays cover the biological events protein–protein binding, protein cleavage, protein function, and enzyme activity/cleavage and are the same ones shown to have an overall median rank above 65000 (**Figure 15**). These results suggest that expansions in chemical and biological space are unable to effectively retrieve the most active compounds for these assay types after the first few iterations.

Chemical Diversity Analysis of Iterative Screening Results. In addition to the rank distribution of the iteratively selected compounds, we also analyzed the percentage of highly active scaffolds cumulatively retrieved. Highly active scaffolds were separately defined for each assay as the Murcko scaffolds[29] belonging to the top 0.5% most active molecules in the assay. While Murcko

scaffolds are useful for assessing structural diversity of cyclic compounds (the definition by Bemis and Murcko[29] is based on ring systems and linkers), this measure of diversity is biased for assays where many aliphatic compounds are hits. In the absence of a more inclusive and/or appropriate definition of scaffold, the following analysis only includes chemical matter with a defined Murcko scaffold.

The average retrieval rate of highly active scaffolds after 10 iterations across all assay types is 41% (~1600 unique scaffolds per assay, ~9 analogs per scaffold), with an average of 14959 compounds screened across all iterations per assay. These results indicate that our method is able to prioritize diverse chemical matter despite much smaller screening sets. In addition, it performs substantially better than a traditional similarity search as the retrieval of highly active scaffolds is only 11% in the first iteration where the similarity search would stop, compared to 41% after 10 rounds of iterative screening.

The percentage of cumulatively retrieved highly active scaffolds steadily increases with the iteration count (**Figure 18**), with the steepest increases occurring in the earliest iterations.

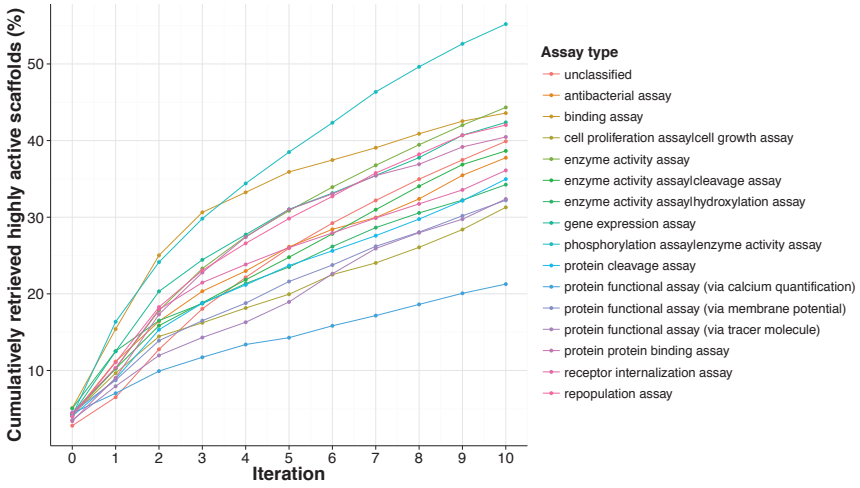


Figure 18. Cumulatively retrieved highly active scaffolds (%). For all assay types, the percentage of cumulatively retrieved highly active scaffolds (scaffolds of the 0.5% most active compounds of the full HTS) steadily increases, with the steepest increases occurring in the earliest iterations. Most assay types display a scaffold retrieval of between 30 and 45% after 10 iterations. The calcium quantification assay showed relatively poor scaffold coverage (~20% after 10 iterations), whereas the phosphorylation assays showed much better scaffold coverage compared to other assay types (~55% after 10 iterations).

Most assay types display a scaffold retrieval of 30–45% after 10 iterations. The calcium quantification assay showed relatively poor scaffold coverage (~20% after 10 iterations), whereas the phosphorylation assay, typically used for kinase inhibitors, showed much better scaffold coverage compared to other assay types (~55% after 10 iterations). Given the presence of many series of high-quality kinase inhibitors from past drug discovery programs in the Novartis screening archive, in combination with the promiscuity of kinase inhibitor binding,[35,36] it is likely that many active inhibitors retrieved are structurally/biologically similar. Hence, this is a possible explanation for the preferred retrieval of a higher number of active scaffolds for phosphorylation assays. Another interesting observation is that the assays for protein–protein binding, protein cleavage, and enzyme activity show mediocre median ranks (> 65000), while having average scaffold retrieval rates (30–40% retrieval after 10 iterations). This suggests that while our iterative screening algorithm (ISA) is able to retrieve many compounds present in the top 0.5% of most active compounds (to an extent comparable with the majority of other assays), many inactive compounds are retrieved as well, resulting in a higher standard deviation in rank (see **Figure 15**). The hypothetically best scaffold retrieval among the top 0.5% of compounds screened would be achieved by sorting the top 0.5% of compounds by activity and picking their scaffolds. We observed that after picking 5000 compounds, this best possible performance retrieves ~75% of highly active scaffolds, compared to ~10–25% of highly active scaffolds (depending on assay type) retrieved iteratively and ~0.4% that would be retrieved if selection was random. In other words, iterative screening of ~15000 compounds recovers a third of the structural diversity of the top 5000 compounds of a 1.3 M compound screen.

The fraction of highly active scaffolds retrieved was also analyzed across all assay types. Here, we determined the fraction of highly active scaffolds for each iteration (see **Figure 19**).

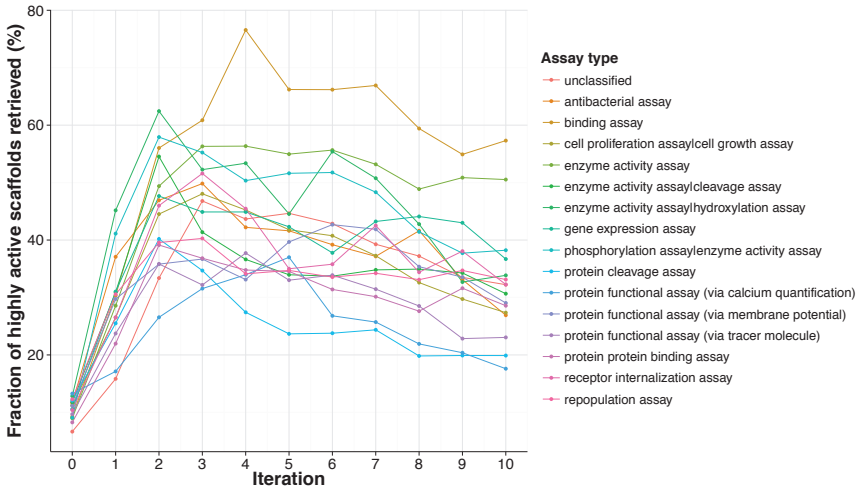


Figure 19. Fraction of highly active scaffolds retrieved (%). The iterative screening algorithm (ISA) exhibits a general trend for all assays: for the first two or three iterations, the fraction of highly active scaffolds retrieved per iteration sharply increases from ~10% to 30–80% depending on assay type (the active scaffolds which are easy to identify are quickly retrieved), after which it slowly decreases, as it becomes increasingly difficult to find the remaining highly active scaffolds. Nevertheless, active scaffolds are still retrieved at the last iterations.

We observed that, in general, the active scaffolds which are easily identified are quickly retrieved: for the first few iterations, the fraction of highly active scaffolds retrieved sharply increases from ~10% to 30–80%, after which it slowly decreases, indicating the progressive difficulty in finding the remaining highly active scaffolds. A possible explanation is the presence of unreachable singletons in the screening archive that are beyond the expansions we implemented thus far.

Visualization of Stepwise Exploration of Chemical Space. In order to illustrate the iterative compound selection in more detail, we showed the expansions for an inhibitory cell-free kinase assay in a network graph (see **Figure 20**).

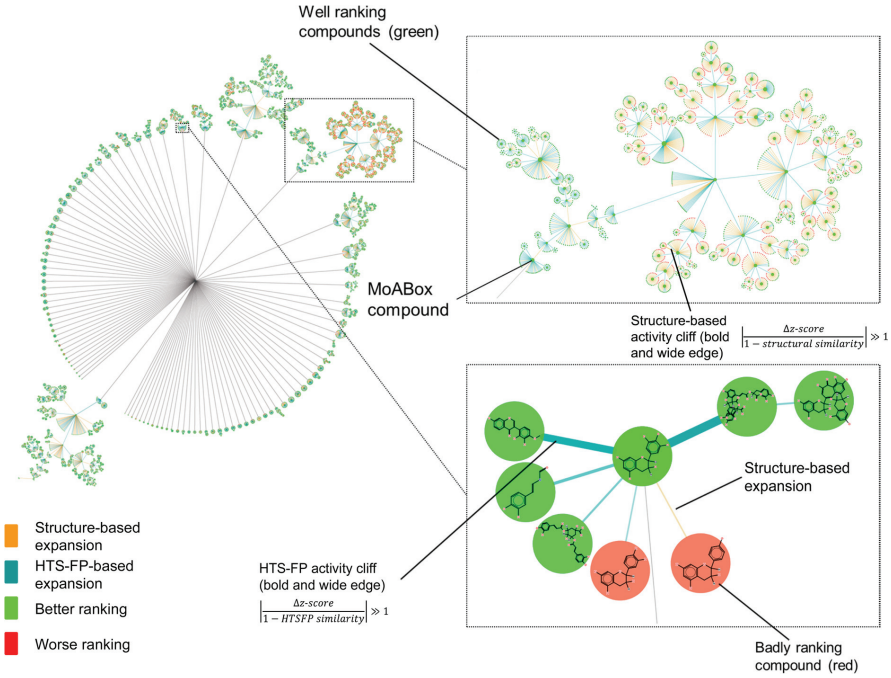


Figure 20. Visualization of stepwise exploration of chemical space for an inhibitory cell-free kinase assay. Expansions for an inhibitory cell-free kinase assay are shown in a network graph. All compounds from the starting set (zeroth iteration) leading to no further expansions have been omitted from the network graph, whereas those that led to at least one further expansion are depicted on the large circle on the left part of the figure. All the compounds present in the subnetwork in the upper-right corner of the figure represent expansions from one single compound from the starting set (MoABox). In the lower-right corner of the figure, we show an example of scaffold hopping, which is commonly caused by expansions based on biological similarity (HTS-FP), enabling the method to explore chemical space that is not reachable via expansions based on chemical similarity. In addition, the depiction of activity cliffs[37] (represented by bold and wide edges) allows the identification of scaffold hopping leading to relatively sharp increases in activity.

All compounds from the starting set (zeroth iteration) leading to no further expansions have been omitted from the network graph, whereas those that lead to at least one further expansion are depicted on the large circle on the left part of the figure. Compounds are color-coded according to their rank (lower/ better and higher/worse ranks are represented by green and red nodes, respectively), and edges are colored according to the expansion type (chemical similarity expansions are orange and biological similarity expansions are turquoise). Certain compounds from the starting set lead to very few further expansions, and hence produce very few branches. Other compounds lead to a larger number of expansions, as can be seen in the upper-right corner of the figure: all the compounds present in that

subnetwork represent expansions from one single compound of the starting set. In the lower-right corner of the figure, we show an example of scaffold hopping, which is commonly observed for biological similarity (HTS-FP) expansions, enabling the method to explore chemical space that is not reachable *via* chemical similarity. In addition, the depiction of activity cliffs[37] (represented by bold and wide edges) allows the identification of scaffold hopping indicative of a relatively sharp increase in activity.

Tuning Iterative Screening to Assay Requirements. The number of compounds triaged per iteration has a large effect: as more compounds are carried forward, both the median ranks and the scaffold retrieval for compounds selected in iterations 1−10 increase (comparison of experiments 2 and 3 with reference, see **Figure 21** and **Figure 22**).

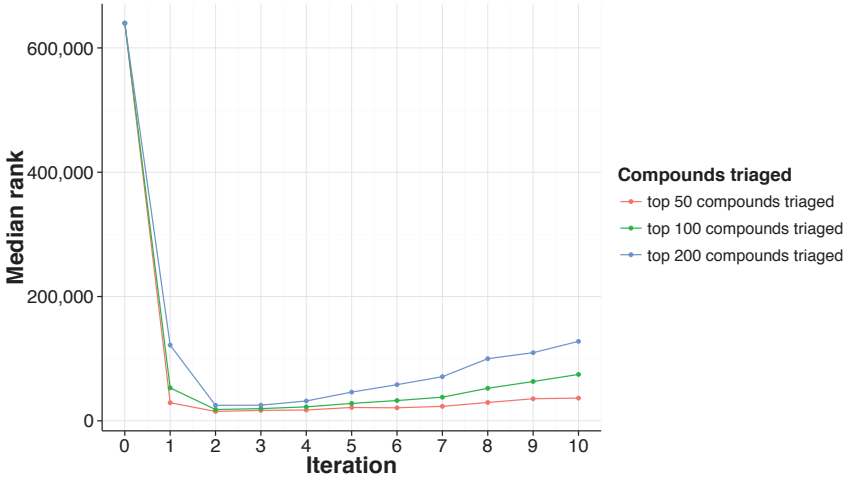


Figure 21. Effect of varying the number of compounds triaged per iteration in terms of median compound rank. As the number of compounds triaged increases, the median ranks consistently increase for iterations 1 to 10. These results are in accordance with our expectations: as the number of triaged compounds is increased (i.e., a less stringent selection criterion is applied for compound triaging), more expansions take place and more compounds are screened overall.

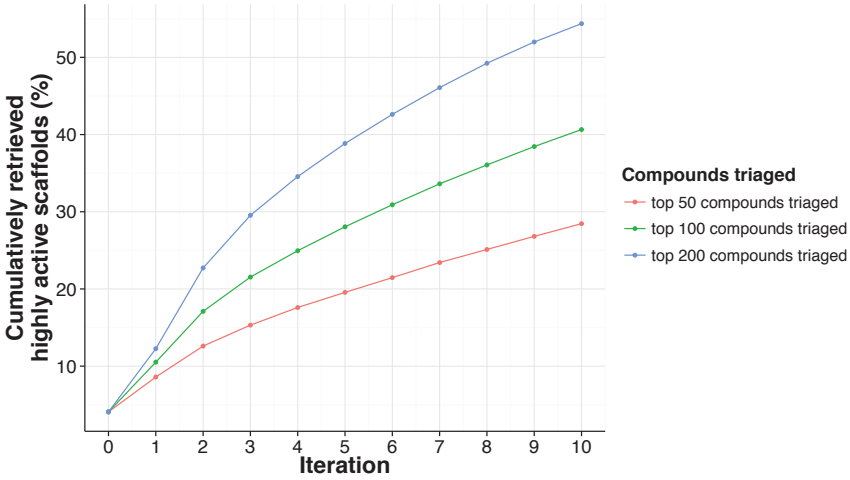


Figure 22. Effect of varying the number of compounds triaged per iteration in terms of percentage cumulatively retrieved highly active scaffolds. As the number of compounds triaged increases, scaffold retrieval is higher as well. These results are in accordance with our expectations: as the number of triaged compounds is increased (i.e., a less stringent selection criterion is applied for compound triaging), more expansions take place and more compounds are screened overall.

When the number of compounds triaged was increased from 50 to 100 and from 100 to 200, median ranks of the compounds selected in iterations 1–10 increased significantly from 23517 to 36101 in the first case and from 36101 to 63721 in the second case (paired *t*–test *p*-values of 1.2×10^{-3} and 3.4×10^{-4}, respectively). Scaffold retrieval increased from 20% to 28% and from 28% to 38% for the same comparisons, with respective paired *t*–test *p*-values of 1.1×10^{-5} and 9.9×10^{-6}. Less stringent hit selection during triaging leads to more subsequent expansions and increases the total number of compounds screened. The overall net result is an increased retrieval of active scaffolds at the cost of screening more inactive compounds as evidenced by higher median ranks.

When investigating the dependence of scaffold coverage on fingerprint type, we found that HTS-FP-based and structure-based expansions accounted for 90% and 50%, respectively, of total highly active scaffold retrieval after 10 iterations. Since HTS-FPs capture the biological profile of compounds, HTS-FP similarity leads to more structurally diverse sets of biologically similar compounds compared to structure-based expansions.

Increasing the Tanimoto[30] cutoff from 0.6 to 0.8 (comparison of experiment 5 to the reference experiment) for structure-based expansions decreased both

median compound ranks from 36101 to 16831 (paired *t*–test *p*-value of 9.4×10^{-4}) and scaffold retrieval from 28% to 16% (paired *t*–test *p*-value of 2.6×10^{-6}). The maximum number of compounds triaged per parent compound did not have a clear effect on the diversity nor the ranks of the compounds screened. Lowering this number from 5 (reference experiment) to 2 (experiment 8) resulted in a 2% higher scaffold retrieval (paired *t*–test *p*-value of 0.047), whereas an increase to 10 (experiment 9) had no significant effect on either median ranks or scaffold retrieval. In summary, the number of compounds triaged was the most influential factor, which can be adjusted depending on the number of compounds one intends (or can afford) to screen.

Finally, iterative screening was repeated with 10 randomly chosen starting sets, and the results were compared to those obtained with the MoABox as a starting set. The latter resulted in better median ranks only until the first iteration, virtually identical median ranks from iteration two onward, and slightly higher scaffold retrieval throughout all iterations. While minor differences across starting sets can be observed, the key findings presented in this study are independent of the precise composition of the starting set. However, the availability of a high-quality starting set, as the MoABox for us, can provide biological insight early on through comprehensive compound annotations.

Conclusions

Even though alluded to in the literature and theoretically appealing, no comprehensive practical evaluation of iterative screening was published. In this study, we have performed an unequaled large-scale validation of iterative screening on 34 HTS assays comprising at least 1300000 compounds and showed greatly improved efficiency over conventional HTS campaigns. For most assays, half of the compounds found by iterative screening of only 1% (~15000 compounds) of the entire collection correspond to the top 5% of the full collection screen. Put differently, screening only 1% of the collection provides ~7500 top-quality hits for further optimization. On average, the compounds selected covered over 40% of the scaffolds belonging to the top 0.5% most active compounds for each assay, hence also ensuring structural diversity. Our method allows for exit points during the iterative screening process: performing large numbers of iterations is not necessary in order to retrieve active compounds, as they are retrieved starting from the first

iteration already, and therefore, a large investment in resources upfront is not required. As expected, the method in its current state performs better for cell-free assays compared to cell-based assays; a future improvement can gear toward physicochemical properties more adapted to cell-based screens.

We used network graphs to visualize the compound selection process, and to highlight activity cliffs,[37] scaffold hopping, and the effect of changing the number of compounds triaged (which was found to have the largest influence on compound selection). As an outlook for further refinement of our method, we propose (1) investigating activity cliffs[37] (to be able to prioritize expansion types) and (2) employing iteratively retrained machine-learning methods[20] to rank the screening collection in parallel to the structure-based and HTS-FP-based expansions currently performed. We believe that the iterative method developed here can easily be fine-tuned for specific assay types, provides multiple exit points, and can potentially lead to considerable savings in both time and resources.

References

(1) Drews, J. (2000) Drug Discovery: A Historical Perspective. *Science (Washington, DC, U. S.) 287*, 1960−1964.

(2) Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today 11*, 277−279.

(3) Mayr, L. M., and Fuerst, P. (2008) The future of high-throughput screening. *J. Biomol. Screening 13*, 443−448.

(4) Phatak, S. S., Stephan, C. C., and Cavasotto, C. N. (2009) High-throughput and in silico screenings in drug discovery. *Expert Opin. Drug Discovery 4*, 947−959.

(5) Mayr, L. M., and Bojanic, D. (2009) Novel trends in high-throughput screening. Curr. Opin. *Pharmacol. 9*, 580−588.

(6) Valler, M. J., and Green, D. (2000) Diversity screening versus focussed screening in drug discovery. *Drug Discov. Today 5*, 286− 293.

(7) Fox, S., Farr-Jones, S., Sopchak, L., Boggs, A., Nicely, H. W., Khoury, R., and Biros, M. (2006) High-Throughput Screening: Update on Practices and Success. *J. Biomol. Screening 11*, 864−869.

(8) Terstappen, G. C., Schlüpen, C., Raggiaschi, R., and Gaviraghi, G. (2007) Target deconvolution strategies in drug discovery. *Nat. Rev. Drug Discovery 6*, 891−903.

(9) Astashkina, A., Mann, B., and Grainger, D. W. (2012) A critical evaluation of in vitro cell culture models for high-throughput drug screening and toxicity. *Pharmacol. Ther. 134*, 82−106.

(10) Huggins, D. J., Venkitaraman, A. R., and Spring, D. R. (2011) Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Biol. 6*, 208−217.

(11) Perez, J. J. (2005) Managing molecular diversity. *Chem. Soc. Rev. 34*, 143−152.

(12) Petrone, P. M., Wassermann, A. M., Lounkine, E., Kutchukian, P., Simms, B., Jenkins, J., Selzer, P., and Glick, M. (2013) Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discov. Today* 18, 674−680.

(13) Willett, P. (1999) Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds. *J. Comput. Biol. 6*, 447−457.

(14) Koutsoukas, A., Paricharak, S., Galloway, W. R. J. D., Spring, D. R., IJzerman, A. P., Glen, R. C., Marcus, D., and Bender, A. (2013) How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J. Chem. Inf. Model. 54*, 230−242.
(15) Baell, J., and Walters, M. A. (2014) Chemical con artists foil drug discovery. *Nature 513*, 481−483.
(16) Che, J., King, F. J., Zhou, B., and Zhou, Y. (2012) Chemical and Biological Properties of Frequent Screening Hits. *J. Chem. Inf. Model. 52*, 913−926.
(17) Stanton, D. T., Morris, T. W., Roychoudhury, S., and Parker, C. N. (1999) Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Model. 39*, 21−27.
(18) Crisman, T. J., Jenkins, J. L., Parker, C. N., Hill, W. A. G., Bender, A., Deng, Z., Nettles, J. H., Davies, J. W., and Glick, M. (2007) Plate cherry picking: a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J. Biomol. Screening 12*, 320−327.
(19) O'Boyle, N. M., Bostrom, J., Sayle, R. A., and Gill, A. (2014) Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem. 57*, 2704−2713.
(20) 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.
(21) 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.
(22) Sams-Dodd, F. (2005) Target-based drug discovery: Is something wrong? *Drug Discov. Today 10*, 139−147.
(23) 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.
(24) Dančík, V., Carrel, H., Bodycombe, N. E., Seiler, K. P., Fomina-Yadlin, D., Kubicek, S. T., Hartwell, K., Shamji, A. F., Wagner, B. K., and Clemons, P. A. (2014) Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screening 19*, 771−781.
(25) Keenan, S. M., Geyer, J. A., Welsh, W. J., Prigge, S. T., and Waters, N. C. (2005) Rational inhibitor design and iterative screening in the identification of selective plasmodial cyclin dependent kinase inhibitors. *Comb. Chem. High Throughput Screening 8*, 27−38.
(26) Schneider, G., Hartenfeller, M., Reutlinger, M., Tanrikulu, Y., Proschak, E., and Schneider, P. (2009) Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol. 27*, 18−26.
(27) RDKit: cheminformatics and machine learning software. http://www.rdkit.org/ (Accessed 2013).
(28) Rogers, D., and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model. 50*, 742−754.
(29) Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem. 39*, 2887−2893.
(30) Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical Similarity Searching. *J. Chem. Inf. Model. 38*, 983−996.
(31) Indigo toolkit, version 1.1.12; GGA Software Services: Cambridge, MA, 2013.
(32) TIBCO Spotfire, version 4.0.4.38; TIBCO Software Inc.: Dublin, OH, 2014.
(33) Dessau, R. B., and Pipper, C. B. (2008) R−project for statistical computing. *Ugeskr. Laeger. 170*, 328−330

(34) Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., Lotia, S., Pico, A. R., Bader, G. D., and Ideker, T. (2012) A travel guide to Cytoscape plugins. *Nat. Methods 9*, 1069−1076.
(35) Paricharak, S., Klenka, T., Augustin, M., Patel, U. A., and Bender, A. (2013) Are phylogenetic trees suitable for chemogenomics analyses of bioactivity data sets: the importance of shared active compounds and choosing a suitable data embedding method, as exemplified on Kinases. *J. Cheminf. 5*, 49−68.
(36) Hanks, S. K., and Hunter, T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J. 9*, 576−596.
(37) Guha, R., and Van Drie, J. H. (2008) Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model. 48*, 646−658.

Chapter five

Data-driven Derivation of an "Informer Compound Set" for Improved Selection of Active Compounds in High-Throughput Screening (manuscript submitted)

Shardul Paricharak, Adriaan P. IJzerman, Jeremy L. Jenkins, Andreas Bender, and Florian Nigsch

Abstract

Despite the usefulness of high-throughput screening in drug discovery, for some systems, low assay throughput or high screening cost can prohibit the screening of large numbers of compounds. In such cases, iterative cycles of screening involving active learning (AL) are employed, creating the need for smaller "informer sets" that can be routinely screened to build predictive models for selecting compounds from the screening collection for follow-up screens. Here, we present a data-driven derivation of an informer compound set with improved predictivity of active compounds in HTS, and validate its benefit over randomly selected training sets on 46 PubChem assays comprising at least 300000 compounds and covering a wide range of assay biology. The informer compound set showed improvement in BEDROC(α=100), PRAUC and ROCAUC values averaged over all assays of 0.015, 0.010 and 0.016, respectively, compared to randomly selected training sets, all with paired *t*–test *p*-values < 10^{-15}. A per-assay assessment showed that the BEDROC(α=100), which is of particular relevance for early retrieval of actives, improved for 40 out of 46 assays, increasing the success rate of smaller follow-up screens. Overall, we showed that an informer set derived from historical HTS activity data can be employed for routine small-scale exploratory screening in an assay-agnostic fashion. This approach led to a consistent improvement in hit rates in follow up screens without compromising on scaffold retrieval. The informer set is adjustable in size depending on the number of compounds one intends to screen, as performance gains are realized for sets with more than 3000 compounds, and this set is therefore applicable to a variety of situations. Finally, our results indicate that random sampling may not adequately cover descriptor space, drawing attention to the importance of the composition of the training set for predicting actives.

Introduction

Over the past three decades, high-throughput screening (HTS) has become a well-established method used during early drug discovery.[1-7] However, low assay throughput or high screening cost can at times prohibit the screening of large numbers of compounds.[8,9] Given this drawback, iterative cycles of design-screen-refine involving active learning (AL) strategies can be used when only a small number of compounds can or should be screened.[10-12] This, in combination with recent advances in machine learning, has recently

prompted efforts to improve bioactivity modeling in order to identify active compounds *in silico*, with the aim of increasing the hit rates in compound screens.[11]

For this purpose, a high-throughput screening fingerprint (HTS-FP) was developed by Petrone *et al.*[13] and later by Dančík *et al.*,[14] which profiles compounds according to their bioactivity across a range of HTS assays. This work was based on the idea that such fingerprints are predictive of compound affinity on targets *not* covered in the fingerprint and showed the value of HTS-FP for virtual screening and biodiverse selection of actives. This concept has previously been explored computationally on smaller datasets,[15-18] but without large-scale experimental validation. More recently, Riniker *et al.*[19] benchmarked the predictive performance of chemical fingerprints and HTS-FP in conjunction with a variety of classification methods across a large number of assays performed in Novartis and those in the public domain (available in PubChem).[20] It was found that random forest (RF) methods with HTS-FP often outperformed machine learning methods developed on chemical descriptors.[19] On a related note, Maciejewski *et al.*[21] explored an experimental design strategy where AL was used to enhance the chemical diversity of large training sets comprising over 50000 compounds, leading to improvement in model performance. While the mentioned studies addressed the dependence of the model on descriptor and classification method used, a comprehensive assessment of how the composition of the initially screened compound set (training set) affects model performance and early retrieval of actives from the remaining screening collection was not performed.

The effectiveness of HTS screening sets in identifying actives has been widely discussed.[22] Given the possible existence of over 10^{63} drug-like molecules,[7] it is remarkable that HTS campaigns comprising "only" 10^6 compounds succeed in finding hits at all.[22-24] A plausible explanation for this is that screening libraries are not random, but rather biased towards biogenic compounds, likely to interact with the druggable proteome. This claim has been reinforced by studies showing the chemical similarity between metabolite space, natural product space and bioactive space.[25-27] A comprehensive analysis by Klekota *et al.*[28] showed that certain "privileged" chemical substructures, such as benzodiazepines,[29] enrich for bioactivity, creating further avenues for modeling the likelihood of compounds being bioactive in *any* therapeutically relevant setting (hereafter referred to as joint bioactivity modeling), rather

than target- or phenotype-specific bioactivity modeling (also shown by Gillet *et al.*).[30]

In this study, we harnessed bioactivity information from a large number of PubChem[20] HTS assays to derive an assay-agnostic "informer compound set" that, once screened, predicts bioactivity better than randomly selected sets for almost all HTS assays, improving the efficiency of subsequent screens. We used AL to iteratively derive this set. Due to the difficulty in implementing AL under extreme class imbalance[31] as is the case for all HTS assays analyzed in his study, activities from multiple assays were combined to derive binary labels representing assay-agnostic bioactivity for each compound. This was based on the idea of joint bioactivity modeling[28,30] and led to a class-balanced dataset suitable for AL. HTS-FPs were used as descriptors, as they showed improved performance over chemical fingerprints.[19] Moreover, this informer set was constructed with the aim to facilitate routine screens, as pre-composed sets are easier to screen routinely from an infrastructure point of view.

Methods

HTS Data. The public HTS data used by Riniker *et al.*[19] was used in this study (see **Tables S1** and **S2** of this reference for the list of assays used). HTS data from the NIH molecular libraries program (MLP) comprising at least 300000 compounds per assay, and submitted by the NCGC, the Scripps Research Institute Molecular Screening Center, or the Burnham Center for Chemical Genomics were extracted from PubChem.[20] This resulted in a total of 141 cell-based and target-based assays (mainly using fluorescence readout technologies), covering a wide range of assay biology (kinases, proteases, ion channels, GPCRs and other target classes). Assay-specific z-scores were calculated for all compounds tested based on the activity measurement used to define the PubChem activity outcome. The set of assays was subsequently split into 2 groups: 95 "group 1 assays" (comprising over 338000 compounds) and 46 "group 2 assays" (comprising 300000–338000 tested compounds, depending on operational turnover of the compound collection at the screening centers). Group 1 assays (referred to as "historical assays" by Riniker *et al.*)[19] were used exclusively for the construction of HTS-FP,[13] a fingerprint used as a descriptor for machine learning, profiling the activity of a compound across HTS assays based on z-scores (float version).[13] Group 2 assays (referred to as "test assays" by Riniker *et al.*)[19] were used for deriving

labels and for model training and testing. This distinction between assay groups ensured that there was no overlap in targets between the two groups.[19]

HTS-FP. For each compound, an HTS-FP was computed, in which each element corresponds to the z-score (based on activity) of the compound in one of the group 1 assays. Missing z-scores (15% of all data points; not every compound is tested in each assay) were assumed to be 0 (the mean of z-scores), as implemented earlier by Riniker *et al.*[19]

Workflow. In this study we tested the performance of bioactivity models developed on an informer set derived with AL. First, we evaluated the performance for predicting bioactivity independent of tested assay (**Figure 23**, joint bioactivity modeling).

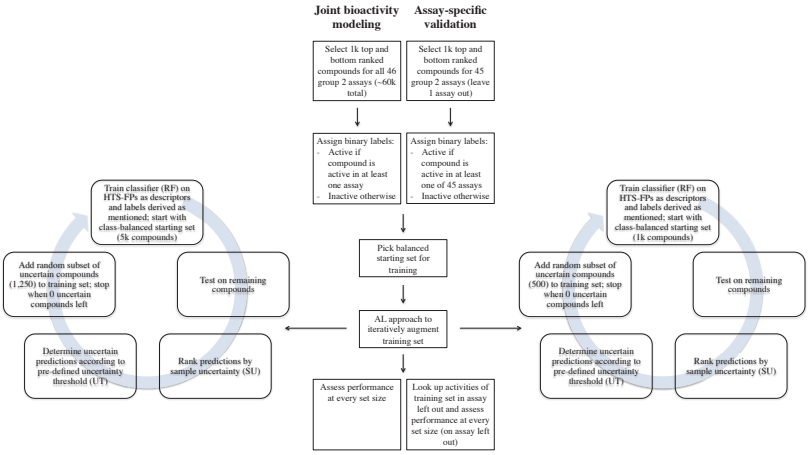


Figure 23. Overview of workflow. In this study, two analyses were performed. Firstly (left), a joint bioactivity model was developed on the 1000 top and bottom ranked (based on z-scores) compounds. An AL approach was used to iteratively augment the training set, for which model performance (ROCAUC) was assessed at every set size. The second analysis (right) involved an assay-specific validation, where a joint bioactivity model was developed on all assays except the assay left out of training. The training set was iteratively augmented with uncertain samples using AL, and at every set size, activities of these compounds were looked up in the assay left out. Subsequently, model performance (ROCAUC, PRAUC, BEDROC) for the training set was assessed on the assay left out, rather than on the joint activities dataset.

Here, activities from group 2 assays were combined to derive binary labels representing assay-agnostic bioactivity for each compound in order to construct a class-balanced dataset suitable for AL. Improved model performance at this step was considered a prerequisite for the more challenging task of predicting actives for individual assays. An assay-specific validation was performed to address the latter task: the informer set was

derived from activity data from 45 group 2 assays and predictivity was assessed on the one assay remaining (**Figure 23**, assay-specific validation). This was repeated 46 times, effectively leaving each group 2 assay out once.

Joint Bioactivity Modeling. The 1000 least and most active compounds (based on z-scores) were selected from each group 2 assay, resulting in a total of 58768 compounds. A skewed distribution of the number of assays these compounds were active in was observed, with 45%, 33%, 12% and 10% of compounds active in 0, 1, 2 and more than 2 assays, respectively. Each compound was labeled as "active" if it was active in *any* of the group 2 assays (as defined by the PubChem activity outcome) or "inactive" otherwise, resulting in a total of 32171 actives and 26597 inactives. This labeling was based on the concept of considering activities independent of the assay they were tested in (joint bioactivity). An RF model (scikit-learn)[32-34] was developed on a randomly selected class-balanced training set of 5000 compounds, and the performance of the model was assessed on the remaining compounds. Using AL, this training set was iteratively augmented with up to 1250 uncertain samples at each iteration, with the aim to improve model performance on the remaining compounds (see "Active Learning" section for more details). The model for this informer set was benchmarked against a model developed on a randomly selected set at each set size using the area under the receiver operating characteristic curve (ROCAUC).

Assay-specific Validation. Here, the informer set was derived from activity data from 45 group 2 assays, and a model was trained on group 1 assay HTS-FPs and labels derived from the one assay left out. The starting set was a randomly selected class-balanced set of 1000 compounds, which was iteratively augmented by up to 500 compounds using AL (see "Active Learning" section for more details). The size of the training and augmentation set was kept smaller here than for the joint bioactivity modeling due to observed improvement in performance at the earlier stages of the algorithm. Performance on the assay left out was assessed at each set size using the ROCAUC, the area under the precision-recall curve (PRAUC),[35] Boltzmann-enhanced discrimination of ROC (BEDROC) (α=100),[36,37] and the retrieval of Murcko scaffolds[38] belonging to the active compounds. The BEDROC(α=100) was used due to its relevance in early retrieval of actives in imbalanced datasets and the PRAUC was used because it captures the effect of the large number of inactive compounds on the model's performance.[35] Both these metrics were therefore considered more relevant than the ROCAUC for the

assay-specific validation (by contrast, for the joint bioactivity modeling the ROCAUC was considered an adequate metric due to class balance).

The model was benchmarked against models developed on a randomly selected set and a set comprising compounds with the highest median z-scores across the 45 assays left in (the frequent hitter set). The latter comparison was included to ensure that the performance gain for the informer set was better than when simply more actives from other assays (including more frequent hitters) were trained on.

Machine Learning. The RF parameters used were: 100 trees (no maximum depth), minimum samples to split = 2, and minimum samples for a leaf = 2.

Active Learning (AL). The AL approach consisted of three iterative steps: (1) training of an RF model, (2) model testing on the remaining compounds and (3) augmenting the training set with a randomly selected subset of uncertain labeled samples (1250 and 500 compounds for the joint bioactivity modeling and assay-specific validation, respectively); when the number of uncertain samples was smaller than the size of the subset, all uncertain samples were selected. The AL algorithm was terminated when the number of uncertain samples was zero. Sample uncertainty (SU) of a given compound c was defined as the absolute probability difference in active versus inactive class predictions:

$$SU_c = \left| p_c^{active} - p_c^{inactive} \right| \quad (4)$$

with SU_c in the range of 0–1 where 0 and 1 represent the most uncertainty and complete certainty in prediction, respectively. Only samples with an SU value smaller than the uncertainty threshold (UT) were considered uncertain. We investigated the effect of varying the UT from 0.5 (least stringent) to 0.01 (most stringent) for the joint bioactivity modeling, and used a UT of 0.1 for the assay-specific validation. The presence of uncertain samples suggests undersampling of bioactivity space. Including these samples could improve model performance over random sampling.[10]

Software Used. The workflow comprised Python scripts for data analysis, using scikit-learn[34] for machine learning and RDKit[39] for scaffold derivation. Tableau[40] was used for data exploration and R[41] was used for the visualization of results.

Results and discussion

The development of an informer set for the prediction of joint bioactivity is presented first (see **Figure 23** – left). Prediction of joint bioactivity allowed the identification of compounds more likely to be bioactive regardless of the assay used. This was followed by a performance assessment of the informer set on individual assays (assay-specific validation; see **Figure 23** – right), and an analysis of scaffold retrieval and set composition. The assay-specific validation was performed in order to determine whether the informer set is more useful than a randomly selected set in predicting actives for novel assays one might perform.

Joint Bioactivity Modeling. The gap in ROCAUC between models developed on the AL sets and on randomly selected sets consistently widens from set sizes of ~5000 onwards (see **Figure 24** – top).

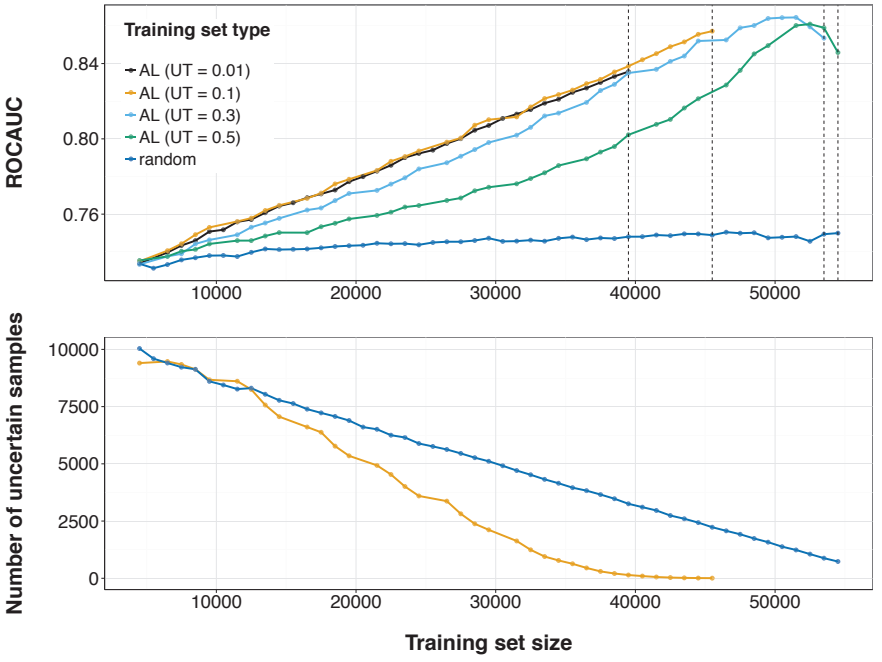


Figure 24. Comparison of model performance for the AL and randomly selected training sets. The ROCAUC (top) is shown for the models trained on AL and randomly selected sets. Performance across all set sizes is consistently better for all AL sets than it is for the randomly selected set. At a set size of 40000 an average gain in performance of 0.08 is observed. In addition, lower *UT* values led to better performance than higher *UT* values. A *UT* value of 0.1 was chosen for the assay-specific validation on the basis of a trade-off between improvement in performance and maximum training set size. For the AL set (*UT* = 0.1), the number of uncertain samples declines faster compared to the randomly selected set (bottom), indicating more efficient sampling of bioactivity space.

At a set size of 40000 an average gain in performance of 0.08 is observed for the AL sets (average ROCAUC of 0.83 compared to 0.75 for randomly selected sets). Stringent *UT* values led to sets with a greater gain in performance at the cost of maximum set size, as fewer samples are classified as uncertain, and the number of uncertain samples reduces to zero earlier in the AL process. Moreover, the number of uncertain samples declines faster for the AL (*UT* = 0.1) set than for the randomly selected set (**Figure 24** – bottom), indicating the benefit of AL in sampling relevant bioactivity space more efficiently. For example, almost all uncertain samples were exhausted for a set size of approximately 40000 using AL, whereas the random set did not exhaust the uncertain samples even at set sizes upwards of 50000. These results indicate that AL is able to consistently sample descriptor space better than random sampling, hereby improving the identification of compounds bioactive in one or more group 2 assays. For further analysis, we chose a *UT* value of 0.1 on the basis of a trade-off between gain in performance and maximum training set size.

Predictive Performance of Informer Set on Individual Assays. In an attempt to translate performance gain in predicting joint bioactivity (see previous section) to performance gain in individual large-scale assays, we performed an assay-specific validation for all group 2 assays. Improved predictive performance in this setting would corroborate the usefulness of an informer set, as no prior information about the assay left out would be required for its construction.

The BEDROC(α=100),[36,37] PRAUC and ROCAUC were calculated for an RF classifier trained on the informer set (AL), a randomly selected set, and the frequent hitter set. These values were averaged over all 46 assay-specific validation experiments and were binned by set size (see **Figure 25**).

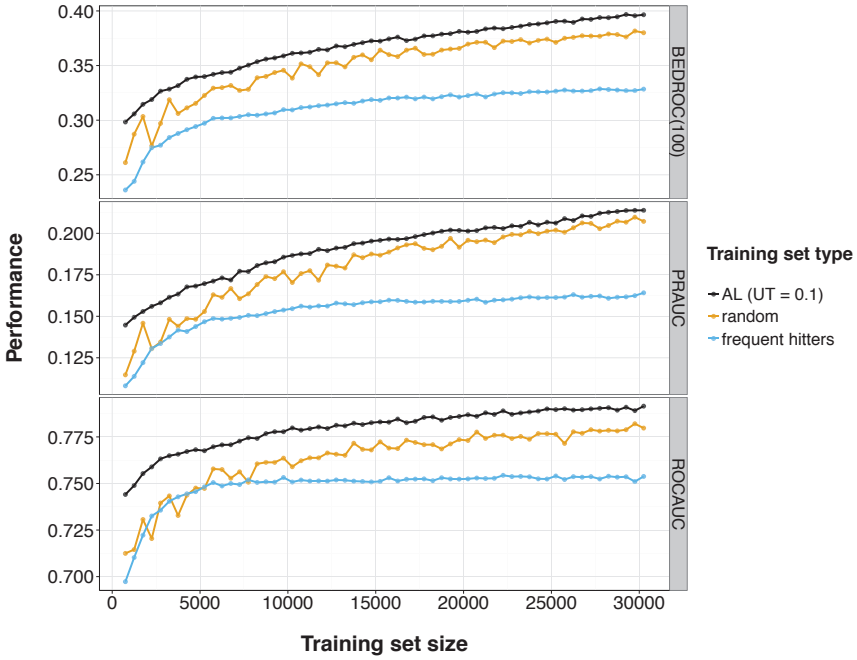


Figure 25. Comparison of model performance for the AL (*UT* = 0.1), random and frequent hitter training sets (assay-specific validation). The BEDROC(α=100)[36,37] (top), PRAUC (middle) and ROCAUC (bottom) binned by set size are shown for all three training sets (bin width=500). The assay-averaged performance for the AL set (all metrics) is consistently better than that for the randomly selected set. For the frequent hitter set, performance is consistently worse than both the AL set and the randomly selected set for training sets larger than 5000 compounds. These results indicate that models trained on the AL set consistently retrieve more actives compared to models trained on the other sets.

The frequent hitter set was used as a benchmark, to ensure that the performance gain of the AL set was better than when simply more actives from other assays (including more frequent hitters) were trained on.

Overall, the performance for the AL set was enhanced compared to the randomly selected set, with an average increase of 0.015, 0.010 and 0.016 in average BEDROC, PRAUC and ROCAUC, respectively (all with paired *t*–test *p*-values < 10^{-15}). The apparent low values of the average BEDROC (0.25-0.40) can be explained by the Boltzmann enhancement, as early retrieval of actives is strongly preferred. Low values of the average PRAUC metric (0.10-0.25) can be explained by the extreme class imbalance: a random classifier would achieve a PRAUC of ~0.007 given the average fraction of actives is only ~0.7%.

For the frequent hitter set, performance is consistently worse for set sizes larger than 5000, indicating that simply including more actives from other

assays does not account for the performance gain observed for the informer set. This finding is in line with the results of the "weak reinforcement strategy" as described in the study by Maciejewski *et al.*[21] Here, training sets with a large number of actives similar in descriptor space (including frequent hitters[42,43] in our study, as the descriptor space is based on bioactivity profiles) were found to be poor at identifying the remaining small number of actives in the test set due to insufficient coverage of descriptor space. By contrast, training sets containing compounds outside the applicability domain, corresponding to uncertain samples in this study, were much better at identifying the remaining actives in the screening collection.

Next, the average improvement in performance over all set sizes of the informer set was calculated separately for each assay (see **Figure 26**).

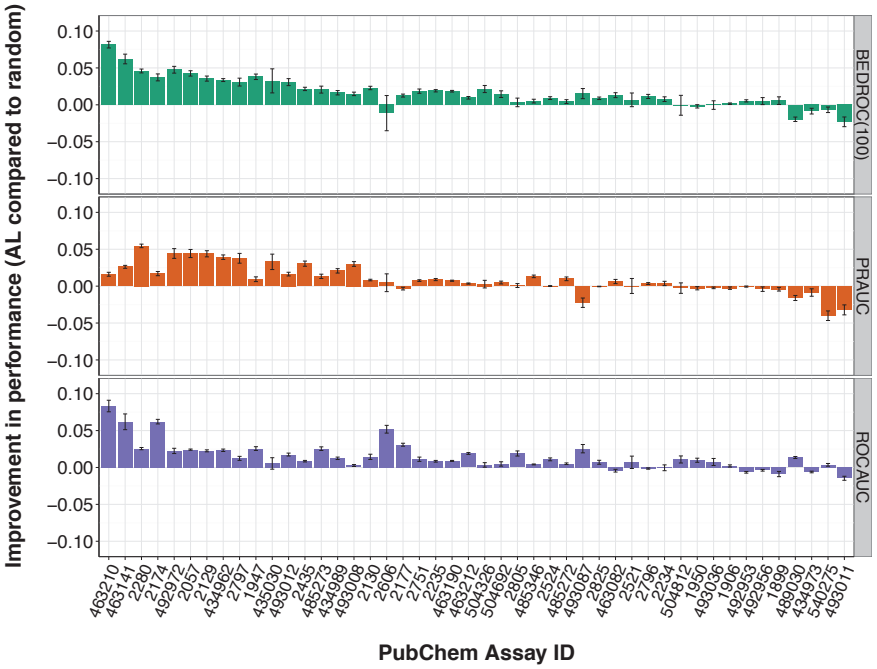


Figure 26. Improvement in model performance for the AL (UT = 0.1) set compared to the randomly selected set for separate assays. The average difference in BEDROC(α=100)[36,37] (top), PRAUC (middle) and ROCAUC (bottom) between the AL set and the randomly selected set is shown for separate assays. Error bars represent standard error of the mean. For 25 out of 46 assays all three metrics improved, whereas the BEDROC(α=100), which is of most relevance for early retrieval of actives,[36,37] improved for 40 out of 46 assays. In practice, the results indicate that if a subsequent screen were performed for each assay, more actives would be retrieved for 40 assays, compared to when random training sets would be used.

For 29 out of 46 assays, all three metrics improved by average 0.02 on average, whereas the BEDROC, which is of most relevance for early retrieval of actives,[36,37] improved for 40 out of 46 assays by 0.02 on average. The best increase in performance was observed for assays number 463210 (caspase 7), 463141 (caspase 3), 2280 (GLD-1 protein) and 2174 (lysophospholipase 1), with BEDROC improvements of 0.08, 0.06, 0.05 and 0.04, respectively. While improvement was modest for most assays, it was consistent, as shown by the error bars representing the standard error of the mean difference in performance between the informer set and the randomly selected set across all sizes. Given the relatively small training sets, varying in size from ~0.3% to 10% of the entire screening collection, large improvements in predictive power over the remaining 90%-99.7% would be unrealistic. We attempted to investigate the cause for the performance loss for the remaining 6 assays, but could not find an explanation: there was no apparent relationship with the average performance for that assay, nor the number of actives in that assay.

Scaffold Retrieval for Individual Assays. We analyzed the scaffold retrieval rate (defined as the retrieved percentage of unique scaffolds belonging to active compounds in the test set; see **Figure 27** – top) and the median z-scores (see **Figure 27** – bottom) of actives identified in the top 5% ranked compounds in order to assess whether these actives were enriched for frequent hitters.

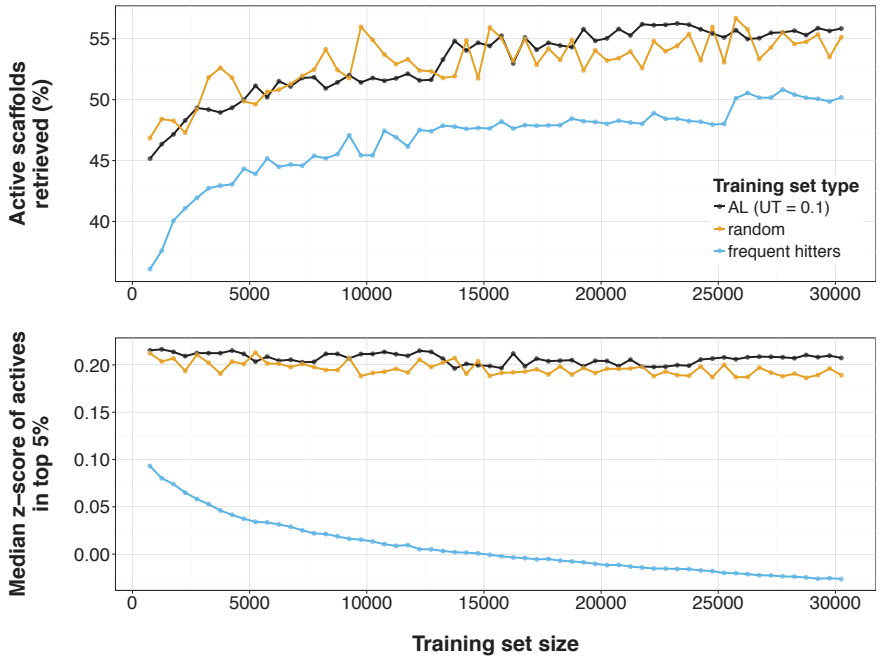


Figure 27. Active scaffold retrieval (%) and median z-scores of actives in top 5% (assay-specific validation). Similar values in scaffold retrieval and the median z-scores of actives in the top 5% ranked compounds for the AL set and the randomly selected set indicate that the AL approach does not compromise on the scaffold retrieval of active compounds, nor does it substantially enrich for frequent hitters. For the frequent hitter set, scaffold retrieval is consistently reduced, hence showing that simply including active compounds from other assays in the training set does not improve the retrieval of diverse sets of actives.

Similar values in scaffold retrieval (45%-55%) and median z-scores (~0.20) for the AL set and the randomly selected set indicate that the AL approach does not compromise on the retrieval of diverse sets of active compounds, nor does it substantially enrich for frequent hitters. Remarkably, the fluctuation in scaffold retrieval is somewhat higher for the random set. This finding is not surprising, as the number of active scaffolds in the training set (which varies in different random sets of compounds due to chance) determines scaffold retrieval in the test set. By contrast, the AL set is iteratively augmented with uncertain compounds that are more likely to have different scaffolds, and hence shows less fluctuation in scaffold retrieval. The frequent hitter set consistently shows worse performance than the other two sets in scaffold retrieval. In addition, the median z-score of the actives retrieved consistently drops from 0.09 to below 0 (**Figure 27** – bottom). The latter drop is likely caused due to fewer compounds with high median z-scores remaining in the

test set as training set size increases. Relative stability of the median z-score is observed for both the AL and random sets, indicating no enrichment for frequent hitters in the training set. In summary, we conclude that when the AL approach is used the scaffold retrieval is not impaired, frequent hitters are not enriched for and at the same time overall hit rates are improved.

Composition of informer set. In order to analyze the composition of the informer set in more detail, we calculated the fraction of the number of active compounds picked from the group 2 assays relative to the number of active compounds for each assay (see **Figure 28**).

Figure 28. Composition of the informer set in terms of active compounds selected from group 2 assays. The heat map represents the composition of the informer set at varying sizes in terms of the fraction of the number of active compounds selected from group 2 assays relative to the number of active compounds for each assay. On the one hand, active compounds from assays number 493008 (troponin C type 1), 2797 (vasopressin V1a receptor), 2435 (oxytocin receptor) and 492972 (platelet-activating factor acetylhydrolase 1b subunit γ) are consistently overrepresented (fold change > 1.3 at a set size of 30000). On the other hand, active compounds from assays number 1950 (EBNA-1 protein), 2805 (intestinal alkaline phosphatase), 493012 (DNA deaminase APOBEC-3G) and 2177 (lysophospholipase 2) are underrepresented (fold change < 0.8 at a set size of 30000). While the AL approach improves performance for all assays, the average BEDROC(α=100) is much higher for the assays with overrepresented actives (0.75) than for the assays with underrepresented actives (0.20).

On the one hand active compounds from assays number 493008 (troponin C type 1), 2797 (vasopressin V1a receptor), 2435 (oxytocin receptor) and 492972 (platelet-activating factor acetylhydrolase 1b subunit γ) are consistently overrepresented in the informer set (maximum fold change > 1.3) while on the other hand active compounds from assays number 1950 (EBNA-1 protein),

2805 (intestinal alkaline phosphatase), 493012 (DNA deaminase APOBEC-3G) and 2177 (lysophospholipase 2) are underrepresented (minimum fold change < 0.8). While the AL approach improves performance for all the assays mentioned above (see **Figure 26**), interestingly, the average BEDROC is much higher for those assays of which the active compounds are *overrepresented* (0.75) than for the assays of which the active compounds are *underrepresented* (0.20). This indicates that more actives are picked from assays already exhibiting good performance. In addition, as determined by Riniker *et al.*[19] the assays of which the active compounds are overrepresented share over 20% of actives with at least six group 1 assays, whereas the assays of which the active compounds are underrepresented share over 20% actives with only at most one group 1 assay (group 1 assays were used to define descriptor space), explaining the difference in BEDROC between these assays.

We attempted to investigate whether bias towards active compounds from particular assays in the informer set was related to improvement in performance over models trained on randomly selected sets for those assays, but could not find any link. We therefore conclude that this improvement in performance is due to better sampling of bioactivity space, as the AL approach iteratively augments the informer set with uncertain samples.

Conclusions

Strategies involving iterative cycles of feedback-driven compound selection and testing can be used when low assay throughput or high screening cost hinders the screening of large compound libraries. This creates the need for the exploratory screening of smaller informer sets to build predictive models for compound selection for follow-up testing. In this study, we performed a data-driven construction of an informer compound set with improved retrieval of actives in a subsequent selection round for apparently unrelated HTS assays. The benefit of this informer set was validated over randomly selected training sets on 46 PubChem[20] assays comprising at least 300000 compounds. Overall, we highlight that such a set – of adjustable size, depending on the number of compounds one intends to screen – can be employed for routine exploratory screening in an assay-agnostic fashion for a gain in predictive power.

Averaged over all assays, an improvement in BEDROC, PRAUC and ROCAUC (of 0.015, 0.010 and 0.016 respectively) was observed with respect to random training sets, all with paired *t*–test *p*-values < 10^{-15}. The informer set

improved the BEDROC for 40 out of 46 assays, indicating better early retrieval of actives. In addition, we found that our approach did not compromise on the retrieval of diverse sets of active compounds, nor did it enrich for frequent hitters, as both scaffold retrieval and the median z-score activity of the actives retrieved were unaffected. The informer set overrepresented actives from certain assays, and underrepresented actives from other assays. Interestingly, while the informer set increased performance for both groups of assays, the BEDROC was much higher (0.75) for the assays of which the actives were overrepresented, than for assays with underrepresented actives (0.20).

We conclude that our AL approach is able to more effectively sample descriptor space, expected to improve the retrieval of active compounds in subsequent screens, thereby reducing the time and expense required to arrive at the same number of hits.

References

(1) Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today 11*, 277–279.

(2) Mayr, L. M., and Fuerst, P. (2008) The future of high-throughput screening. *J. Biomol. Screen. 13*, 443–448.

(3) Phatak, S. S., Stephan, C. C., and Cavasotto, C. N. (2009) High-throughput and in silico screenings in drug discovery. *Expert. Opin. Drug Discov. 4*, 947–959.

(4) Mayr, L. M., and Bojanic, D. (2009) Novel trends in high-throughput screening. *Curr. Opin. Pharmacol. 9*, 580–588.

(5) Valler, M. J., and Green, D. (2000) Diversity screening versus focussed screening in drug discovery. *Drug Discov. Today 5*, 286–293.

(6) Fox, S., Farr-Jones, S., Sopchak, L., Boggs, A., Nicely, H. W., Khoury, R., and Biros, M. (2006) High-Throughput Screening: Update on Practices and Success. *J. Biomol. Screen. 11*, 864–869.

(7) Koutsoukas, A., Paricharak, S., Galloway, W. R. J. D., Spring, D. R., IJzerman, A. P., Glen, R. C., Marcus, D., and Bender, A. (2014) How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model. 54*, 230–242.

(8) Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug. Discov. 1*, 882–894.

(9) Astashkina, A., Mann, B., and Grainger, D. W. (2012) A critical evaluation of in vitro cell culture models for high-throughput drug screening and toxicity. *Pharmacol. Ther. 134*, 82–106.

(10) Settles, B. (2010) Active Learning Literature Survey. *Mach. Learn. 15*, 201–221.

(11) Reker, D., and Schneider, G. (2015) Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today 20*, 458–465.

(12) 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.

(13) 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.
(14) Dančík, V., Carrel, H., Bodycombe, N. E., Seiler, K. P., Fomina-Yadlin, D., Kubicek, S. T., Hartwell, K., Shamji, A. F., Wagner, B. K., and Clemons, P. A. (2014) Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screen. 19*, 771–781.
(15) Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, Å., Bukar, R., Bauer, K. E., Dilley, H., and Rocke, D. M. (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol. 2*, 107–118.
(16) Bender, A., Jenkins, J. L., Glick, M., Deng, Z., Nettles, J. H., and Davies, J. W. (2006) "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. 46*, 2445–2456.
(17) Nguyen, H. P., Koutsoukas, A., Mohd Fauzi, F., Drakakis, G., Maciejewski, M., Glen, R. C., and Bender, A. (2013) Diversity Selection of Compounds Based on "Protein Affinity Fingerprints" Improves Sampling of Bioactive Chemical Space. *Chem. Biol. Drug Des. 82*, 252–266.
(18) Givehchi, A., Bender, A., and Glen, R. C. (2006) Analysis of activity space by fragment fingerprints, 2D descriptors, and multitarget dependent transformation of 2D descriptors. *J. Chem. Inf. Model. 46*, 1078–1083.
(19) 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.
(20) 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.
(21) 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.
(22) Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J., and Shoichet, B. K. (2010) Quantifying Biogenic Bias in Screening Libraries. *Nat. Chem. Biol. 5*, 479–483.
(23) Fox, S., Farr-Jones, S., Sopchak, L., Boggs, A., and Comley, J. (2004) High-throughput screening: searching for higher productivity. *J. Biomol. Screen. 9*, 354–358.
(24) Pereira, D. A., and Williams, J. A. (2007) Origin and evolution of high throughput screening. *Br. J. Pharmacol. 152*, 53–61.
(25) Ertl, P., Roggo, S., and Schuffenhauer, A. (2008) Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model. 48*, 68–74.
(26) Gupta, S., and Aires-de-Sousa, J. (2007) Comparing the chemical spaces of metabolites and available chemicals: Models of metabolite-likeness. *Mol. Divers. 11*, 23–36.
(27) O'Hagan, S., Swainston, N., Handl, J., and Kell, D. B. (2015) A "rule of 0.5" for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics 11*, 323–339.
(28) Klekota, J., and Roth, F. P. (2008) Chemical substructures that enrich for biological activity. *Bioinformatics 24*, 2518–2525.
(29) Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., Lundell, G. F., Veber, D. F., and Anderson, P. S. (1988) Methods for drug discovery: devlopment of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem. 31*, 2235–2246.
(30) Gillet, V. J., Willett, P., and Bradshaw, J. (1997) Identification of biological activity profiles using substructural analysis and genetic algorithm. *J. Chem. Inf. Comput. Sci. 38*, 165–179.

(31) Attenberg, J., and Ertekin, S. (2013) Class imbalance and active learning, in *Imbalanced Learning: Foundations, Algorithms, and Applications, First Edition* (He, H., and Ma, Y., Eds.), pp 101–149. John Wiley & Sons, Inc.
(32) Breiman, L. (2001) Random Forests. *Mach. Learn. 45*, 5–32.
(33) Riniker, S., Fechner, N., and Landrum, G. A. (2013) Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing. *J. Chem. Inf. Model. 53*, 2829–2836.
(34) 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., and Duchesnay, É. (2011) Scikit-learn : Machine Learning in Python. *J. Mach. Learn. Res. 12*, 2825–2830.
(35) Davis, J., and Goadrich, M. (2006) The Relationship Between Precision-Recall and ROC Curves, in *Proceedings of the 23rd International Conference on Machine learning*, pp 233–240.
(36) Truchon, J., and Bayly, C. I. (2007) Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model. 47*, 488–508.
(37) Riniker, S., and Landrum, G. A. (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform. 5*, 26–42.
(38) Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem. 39*, 2887–2893.
(39) RDKit: cheminformatics and machine learning software (http://www.rdkit.org/); 2013.
(40) Tableau Desktop, version 9.0.1; Tableau Software Inc., 2015.
(41) Dessau, R. B., and Pipper, C. B. (2008) R–project for statistical computing. *Ugeskr. Laeger. 170*, 328–330.
(42) Baell, J., and Walters, M. A. (2014) Chemical con artists foil drug discovery. *Nature 513*, 481–483.
(43) Che, J., King, F. J., Zhou, B., and Zhou, Y. (2012) Chemical and Biological Properties of Frequent Screening Hits. *J. Chem. Inf. Model. 52*, 913–926.

Chapter six

Proteochemometric Modelling Coupled to *in Silico* Target Prediction: an Integrated Approach for the Simultaneous Prediction of Polypharmacology and Binding Affinity/Potency of Small Molecules

Reproduced from Shardul Paricharak, Isidro Cortés-Ciriano, Adriaan P. IJzerman, Thérèse E. Malliavin, and Andreas Bender. (2015) *J. Cheminform.* 7, 15-25.

Abstract

The rampant increase of public bioactivity databases has fostered the development of computational chemogenomics methodologies to evaluate potential ligand-target interactions (polypharmacology) both in a qualitative and quantitative way. Bayesian target prediction algorithms predict the probability of an interaction between a compound and a panel of targets, thus assessing compound polypharmacology qualitatively, whereas structure-activity relationship techniques are able to provide quantitative bioactivity predictions. We propose an integrated drug discovery pipeline combining *in silico* target prediction and proteochemometric modelling (PCM) for the respective prediction of compound polypharmacology and potency/affinity. The proposed pipeline was evaluated on the retrospective discovery of *Plasmodium falciparum* DHFR inhibitors. The qualitative *in silico* target prediction model comprised 553084 ligand-target associations (a total of 262174 compounds), covering 3481 protein targets and used protein domain annotations to extrapolate predictions across species. The prediction of bioactivities for plasmodial DHFR led to a recall value of 79% and a precision of 100%, where the latter high value arises from the structural similarity of plasmodial DHFR inhibitors and *T. gondii* DHFR inhibitors in the training set. Quantitative PCM models were then trained on a dataset comprising 20 eukaryotic, protozoan and bacterial DHFR sequences, and 1505 distinct compounds (in total 3099 data points). The most predictive PCM model exhibited $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.79 and 0.59 pIC_{50} units respectively, which was shown to outperform models based exclusively on compound ($R^2_{0\ test}/RMSE_{test}$ = 0.63/0.78) and target information ($R^2_{0\ test}/RMSE_{test}$ = 0.09/1.22), as well as inductive transfer knowledge between targets, with respective $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.76 and 0.63 pIC_{50} units. Finally, both methods were integrated to predict the protein targets and the potency on plasmodial DHFR for the GSK TCAMS dataset, which comprises 13533 compounds displaying strong anti-malarial activity. 534 of those compounds were identified as DHFR inhibitors by the target prediction algorithm, while the PCM algorithm identified 25 compounds, and 23 compounds (predicted $pIC_{50} > 7$) were identified by both methods. Overall, this integrated approach simultaneously provides target and potency/affinity predictions for small molecules.

Introduction

In recent years it has been demonstrated that drugs exert their therapeutic effect by modulating more than one target, in fact six on average.[1] Therefore, the early evaluation of the bioactivity profiles of lead compounds is essential for the success in developing new drugs, although efficacy is sometimes attained by the inhibition of single targets, e.g., viral proteins. Similarly, understanding drug polypharmacology can help in anticipating drug adverse effects.[2]

In parallel, the availability of public bioactivity databases has enabled the application of large-scale chemogenomics techniques to, among others, predict protein targets for small molecules, and to predict their affinity on therapeutically interesting targets.[3] These techniques capitalize on bioactivity data to infer relationships between the compounds, encoded with numerical descriptors, and their targets, which can be represented as labels in a classification model or explicitly encoded by e.g., protein or amino acid descriptors.[4]

In silico target prediction algorithms assess potential compound polypharmacology through the computational evaluation of the (functionally unrelated) targets modulated by a given compound, or its selectivity to species-specific targets, as they predict the probability of interaction of that compound with a panel of targets.[5] Initially, target prediction models were developed using Laplacian-modified Naïve Bayesian classifiers[6] and the Winnow algorithm.[7] Later, Keiser *et al.*[8] developed a model which related biological targets based on ligand similarities and ranked the significance of the resulting similarity scores using the Similarity Ensemble Approach (SEA), followed by Wale and Karypis[9] who applied SVM and ranking perceptron algorithms to rank targets for a given compound. More recently, Koutsoukas *et al.*[10] compared the performance of both the Naïve Bayesian and Parzen-Rosenblatt Window classifiers, concluding that the overall performance of both methods is comparable though differences were found for certain target classes.

The ligand-target prediction methods described above generally predict the likelihood of interaction with a target, and they do not predict compound affinity or potency (e.g., K_i or IC_{50}). On the other hand, quantitative bioactivity prediction techniques, e.g., proteochemometric modeling (PCM),[3] predict the potency or affinity for compound-target pairs, normally in the form of pIC_{50} or pK_i values. PCM combines information from compounds and related

targets, e.g., orthologs, in a single machine learning model,[3,11] which enables the simultaneous modeling of chemical and biological information, and thus the prediction of compound affinity and selectivity across a panel of targets. Nonetheless, the effects of a compound at the cellular or the organism level are poorly understood in this case, as these methods cannot account for the interactions of a compound with other unrelated targets, which are not captured in the PCM model.

Given the limitations of both purely qualitative and purely quantitative bioactivity modeling approaches, in the current work, we propose an integrated drug discovery approach, combining *in silico* target prediction for the qualitative large-scale evaluation of compound bioactivity, and PCM for the quantitative prediction of compound potency. The proposed approach was evaluated on the discovery of DHFR inhibitors for *Plasmodium falciparum* (*P. falciparum*), the causative agent of the most dangerous form of malaria.[12] Whilst there are multiple anti-malarial drugs on the market, resistance to anti-malarial drugs is on the rise,[13,14] and there are only 21 compounds in clinical or pre-clinical trials.[15]

In order to combat the lack of novel drugs for malaria, big pharmaceutical companies have generated a wealth of phenotypic data, namely the GlaxoSmithKline (GSK) TCAMS dataset, as well as the Novartis-GNF Malaria Box.[16,17] Both datasets contain phenotypic readouts, describing how effective the compounds present in the datasets are in inhibiting the growth of *P. falciparum*. Nonetheless, none of them contain annotations about the *P. falciparum* target(s) involved, making it a challenge to elucidate the mode of action (MoA) of the compounds in the dataset, and hence, making the dataset difficult to interpret. This renders these datasets a very suitable case study for the algorithms we are presenting in this work.

In the context of malaria drug discovery, previous studies have applied machine learning algorithms to predict whether plasmodial proteins are secretory proteins based on their residue composition,[18] and to predict the bioactivities of compounds against particular plasmodial targets.[19,20] These approaches, though, did not account for the polypharmacology of anti-malarial compounds.

To overcome the limitations of these methods, we now integrate both *in silico* target prediction and PCM in a unified drug discovery approach. As illustrated in **Figure 29**, the target prediction algorithm used in this study, trained on approximately 553084 bioactivity data points spanning 3481

targets, used a domain-based similarity metric between targets to extrapolate target predictions from one species to another.

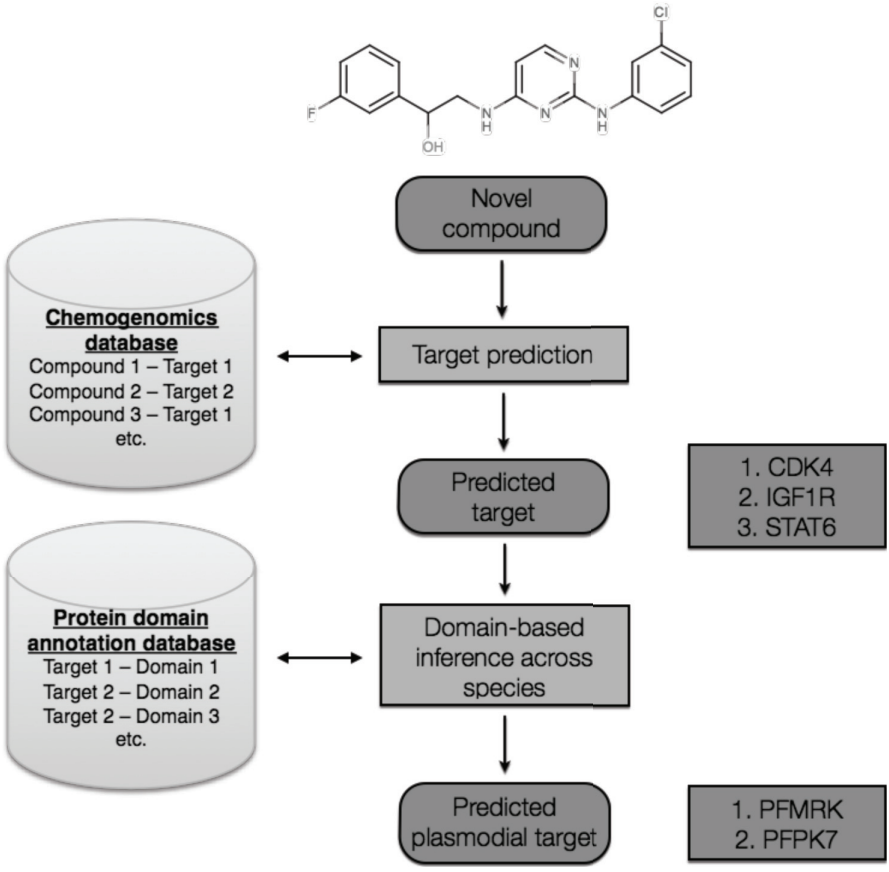


Figure 29. Schematic overview of *in silico* target prediction and domain-based extrapolation workflow. The conventional *in silico* target prediction approach[10] is extended in this study by using protein domain annotations to extrapolate from non-plasmodial target predictions to protein target predictions in *P. falciparum.* This concept is generally applicable across organisms, in particular to those for which little bioactivity data is currently available.

Non-plasmodial targets were then extrapolated to plasmodial targets. Besides, the PCM model was trained on a dataset composed of 20 eukaryotic, protozoan and bacterial DHFR sequences, and of 1505 different DHFR inhibitors and a total of 3099 data points. To exploit the complementarity of the two prediction methods, *in silico* target prediction was used to predict MoA hypotheses for the anti-malarial compounds in the GSK TCAMS phenotypic dataset, whereas PCM was employed to quantify compound potency (pIC_{50}).

Methods

Exploratory Principal Component Analysis (PCA) of PCM and Target Prediction Datasets. A PCA was performed for compounds contained in the PCM dataset, as well as for those annotated on *P. falciparum* and *T. gondii* in the target prediction dataset. The Spearman's rank correlation coefficient was calculated for all pairs of compound descriptors, based on both physicochemical descriptors and Morgan fingerprints, thus defining a square correlation matrix. The PCA analysis was performed on this matrix in order to avoid the direct application of PCA on binary descriptors, i.e., Morgan fingerprints. Visualization was performed using R and Vortex.[21]

Target Prediction.

Training dataset. Bioactivity data were extracted from ChEMBL16[22] according to the protocol described by Koutsoukas *et al.*[10] The extracted data contained approximately 4 million bioactivities covering approximately 8000 biomolecular targets, of which approximately 4000 targets were proteins.[22,23] Compound-target pairs were selected according to the following criteria: (i) K_i, K_d, IC_{50} or EC_{50} bioactivity values equal to or lower than 10 μM, and (ii) targets annotated with a confidence score of 8 (homologous single protein target assigned) or 9 (direct single protein target assigned). Subsequently, ligand structures were processed with the ChemAxon standardizer version 5.12.0,[24] with the following options: "Remove fragment", "Neutralize", "Aromatize", "Clean2D", "Tautomerize" and "Remove explicit hydrogens". After standardization, the entries with ligands annotated against multiple targets were detected based on their canonical SMILES and removed using custom Perl scripts, resulting in a training set of 553084 instances (262174 compounds) covering 3481 protein targets. The bioactivity data of *P. falciparum* (1513 instances – 1379 compounds covering 41 protein targets) was omitted from this dataset for training purposes. InterPro[25] domain annotations were retrieved for all protein targets using the Uniprot database.[26]

P. falciparum dataset. The *P. falciparum* dataset was built using the same criteria as described above, resulting in a set comprising 41 *P. falciparum* targets and 1379 compounds. In addition, all annotated and reviewed *P. falciparum* targets from Uniprot were downloaded, resulting in a total of 148 *P. falciparum* protein targets. Finally, InterPro domain annotations were retrieved for all protein targets using the Uniprot database.

GSK TCAMS dataset. Approximately 2 million compounds present in GSK's screening collection have been tested in vitro by GSK for inhibitors of *P.*

falciparum's intraerythrocytic cycle based on growth inhibition assays.[17] Briefly, assays were performed on both the reference laboratory strain 3D7, as well as on the multidrug resistant strain Dd2, where parasite growth was evaluated using LDH activity.[17] 19451 compounds were identified as primary hits inhibiting the 3D7 strain growth by more than 80% at 2 μM concentration, of which 13533 compounds displayed 80% or higher inhibition of parasite growth in at least 2 of the 3 assay runs in independent follow-up experiments. Hence, these 13533 compounds were considered as confirmed inhibitors (confirmation rate > 70%).

Descriptors. A circular fingerprint implementation, Molprint2D[27,28] was used for encoding molecular structures, since this method has previously been shown to capture structural aspects related to bioactivity better than most other descriptors in comparative studies.[29] This descriptor is based on count vectors of heavy atoms present at a topological distance from each heavy atom of a molecule.[28] For the present study, the pybel implementation was used.[30]

Target prediction algorithm. A multiclass Laplacian-modified Naïve Bayesian classifier, as described by Nigsch *et al.*[7] and later implemented by Koutsoukas *et al.*[10] was implemented to classify the bioactivity dataset and to be able to predict targets for novel compounds. For the query molecule $\boldsymbol{x}$, consisting of a set of n Molprint2D features f_i, the likelihood to be active against a protein target ω_α was calculated using the following equation:

$$S_{\boldsymbol{\omega_\alpha}}(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} log\left(\frac{N_{i,\omega_\alpha}+1}{N_i \times p(\omega_\alpha)+1}\right) + log\left(\frac{\prod_{i=1}^{d} p(f_i)}{p(\boldsymbol{x})}\right) \quad (5)$$

where $S_{\boldsymbol{\omega_\alpha}}(\boldsymbol{x})$ is the logarithmic likelihood score (proportional to the likelihood of bioactivity), N_{i,ω_α} is the total number of occurrences of feature f_i in protein class ω_α and N_i is the total number of occurrences of feature f_i in the entire training set. Furthermore, $p(\omega_\alpha)$ is the prior probability of protein class ω_α. The prior probability quantifies how likely a compound is active against protein target ω_α in the absence of any feature information. It can be calculated as follows:

$$p(\omega_\alpha) = \frac{N_{\omega_\alpha}}{N} \quad (6)$$

where N_{ω_α} is the number of instances (i.e., bioactivities) in class ω_α and N is the total number of instances. The predictive performance of this model was

assessed in terms of average class-specific recall and precision. Only target classes with 20 or more data points in the *P. falciparum* dataset were considered as suitable for testing due to a sufficient number of data points, resulting in a total of 16 target classes.

Domain-based extrapolation to P. falciparum targets. For each analyzed compound, the top *n* ranked predicted targets were compared to all 148 *P. falciparum* targets in terms of their InterPro domain composition. *P. falciparum* targets with an InterPro domain Tanimoto similarity above a variable cutoff were considered as predicted, but were not ranked. The cutoff value varied between 0.5 and 1, where 1 means that only orthologous proteins are considered. The target prediction and domain-based extrapolation pipeline are illustrated in **Figure 29**. The domain extrapolation extends the target prediction approach[10,31] by using InterPro protein domain annotations to extrapolate from predicted non-plasmodial targets to *P. falciparum* targets. This is conceptually similar to a previously reported study for extrapolating bioactivities between species,[32] and its application to *M. tuberculosis*.[33]

The inclusion of plasmodial DHFR (CHEMBL1939) bioactivity data was expected to drastically improve the performance, and this was tested in the following way. A 2-fold cross validation (CV) was performed: the instances annotated on plasmodial DHFR were split into 2 half subsets, where one subset was added to the training set and the other half was used as a test set (and *vice versa*).

Proteochemometric Modeling.

Dataset. IC_{50} values with a confidence score of 8 or 9 for 20 DHFR sequences were retrieved from ChEMBL16[22] and this initial dataset comprised 5827 data points. In the cases where a compound-target combination had more than one annotated bioactivity value, the set of bioactivities was replaced by its mean value. This procedure is robust, because the standard deviation of the differences was smaller than 0.1 pIC_{50} unit in more than 90% of the cases. This resulted in a dataset including 3099 distinct compound-target combinations. The matrix completeness of the dataset, calculated as the number of compound-target combinations present in the dataset over the total number of possible compound-target combinations, was 10.3%. Compounds included in the PCM dataset were not present in the target prediction dataset.

Descriptors. Chemical structures were standardized and cleaned with the function *StandardiseMolecules* of the R package *camb* using the default parameters[34] and PaDEL descriptors (1-D and 2-D). Morgan fingerprints were

calculated in the same environment. The function *AA_Descs* was used to calculate amino acid descriptors (3 Z-scales). To describe the target space, the residues in the binding site of human DHFR (PDB ID: 1OHJ)[35] within a sphere of 10 Å centered around the ligand were selected. The corresponding residues for the other 19 proteins were obtained from a sequence alignment realized with Clustal Omega.[36]

Proteochemometric modeling. All descriptor values were centered to zero mean and scaled to unit variance. The dataset was split into six subsets, five of which were used to train models, and the sixth, test set, was withheld to assess the predictive ability of the models.[37] The hyperparameter values for all PCM models were optimized by 5-fold cross validation.[38] To assess both model predictive ability and performance, the pIC_{50} values for the test set were predicted, thus providing the external validation by calculating $RMSE_{test}$ and $R^2_{0\ test}$ between the observed and the predicted pIC_{50} values:

$$R^2_{0\ test} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i^{r0})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{(y-\hat{y})^2}{N}} \quad (8)$$

where N represents the size of the test set, y_i the observed, $\hat{y}_i$ the predicted, and $\bar{y}$ the average pIC_{50} values of those datapoints included in the test set, and $\hat{y}_i^{r0} = s\hat{y}$, with $s = \sum y_i\hat{y}_i \,/\, \sum\hat{y}_i^2$. Both internal ($RMSE_{int}$ and R^2_{int}) and external validation ($RMSE_{test}$ and $R^2_{0\ ext}$) were assessed according to the criteria proposed by Tropsha *et al.*[39,40] and calculated using the *Validation function* of the R package *camb*.[34]

In order to assess whether the combination of compound and target information in a single PCM model constitutes an advantage with respect to one-space (ligand space and target space) models, two validation scenarios were explored. Firstly, a Family QSAR model[41] was trained exclusively on compound descriptors. High performance of this model is expected in cases where the bioactivities of the same compound on different targets are highly correlated. Secondly, the Family QSAM[41] model was trained on target descriptors only. In this case, high performance would indicate that the activities of a diverse set of compounds are correlated on a panel of targets. Thus, compound activities would largely depend on the target, and to a much lesser extent on the ligand structures.

Additionally, an inductive transfer PCM model (PCM IT) was trained to assess whether the performance of PCM models arises from explicit learning (EL), where the knowledge is extracted from target descriptors, or inductive transfer (IT). In IT the knowledge acquired when predicting compound bioactivities on a given target is exploited to predict the bioactivity of those compounds on another target.[41] In the PCM IT model, targets were described with identity fingerprints (IFP), which are calculated as follows:

$$IFP(i,j) = \delta(i - j)\big(i,j \in 1,\ldots,N_{targets}\big) \quad (9)$$

where δ is the Kronecker delta function and $N_{targets}$ the number of distinct targets. The performance of the models was assessed on a *per* target basis by training ten PCM models, each on a different subset of the whole dataset. Subsequently, $RMSE_{test}$ and $R^2_{0\,test}$ values were calculated on subsets of the test set grouped by target.

Machine learning implementation. Support Vector Machines (SVM),[42] Gradient Boosting Machines (GBM),[43] Gaussian Processes (GP),[44] and Random Forest (RF)[45] models were built with the R package *camb*.[34,46] The target prediction algorithm was implemented in Perl.

Results and discussion

Exploratory Analysis of PCM and Target Prediction Datasets. A PCA (**Figure 30**) was performed for the compounds annotated to be active against plasmodial DHFR and those active against *T. gondii* DHFR.

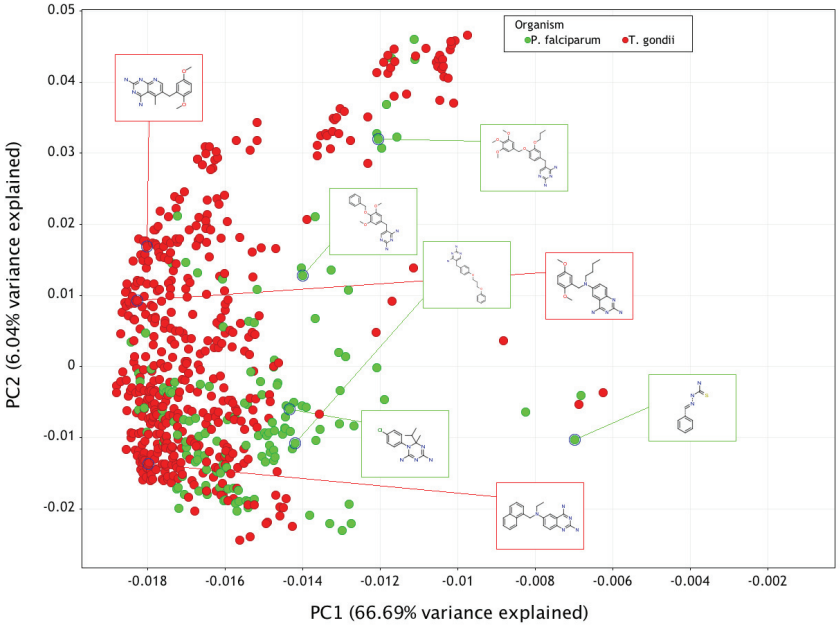


Figure 30. PCA of the compounds annotated as actives against plasmodial DHFR (green) as well as *T. gondii* DHFR (red). Overall, plasmodial DHFR inhibitors cover a substantial portion of the chemical space occupied by *T. gondii* DHFR inhibitors. However, some clusters of *T. gondii* DHFR inhibitors are located in additional chemical space not covered by the plasmodial inhibitors (red boxes). These clusters contain compounds with bicyclic ring systems. By contrast, plasmodial inhibitors only contain unfused rings (green boxes). These observations explain why recall is low (~35%) when plasmodial DHFR inhibitors are excluded from the training set: *T. gondii* inhibitors do not cover all relevant chemical space, particularly the space occupied by compounds with unfused ring systems.

The first two principal components explain 72.73% of the variance. In the two dimensions visualized for the descriptor space used here, the plasmodial inhibitors cover a substantial portion of the chemical space occupied by the *T. gondii* DHFR inhibitors. However, there are still a number of clusters of *T. gondii* DHFR inhibitors that occupy novel space not covered by plasmodial inhibitors. Compounds from these clusters contain bicyclic ring systems (shown in red boxes in **Figure 30**). On the other hand, there are also clusters of plasmodial inhibitors that occupy space not covered by *T. gondii* inhibitors: these plasmodial inhibitors do not contain bicyclic rings, but instead contain unfused rings (5 scaffolds identified shown in green boxes in **Figure 30**). In addition to the previous analysis, a PCA was also performed for the compounds present in the PCM dataset, where the first two principal

components explained 51.77% of the variance. Clusters contain compounds whose bioactivities on several targets are included in the dataset, thus indicating that compounds are overall structurally similar across the 20 DHFR sequences considered.

Application of Target Prediction for MoA prediction. The performance of the target prediction algorithm was assessed for varying values of *n*, which represents the top number of non-plasmodial predictions considered for extrapolation. It can be seen that performance varies widely across target classes: for most targets, including all aminopeptidases, calcium-dependent protein kinase 1, protein kinase Pfmrk, glucose-6-phosphate-1-dehydrogenase, dihydroorotate dehydrogenase, dUTP pyrophosphatase and enoyl-acyl-carrier protein reductase, performance is low, with both recall and precision values below 30%. For a small number of targets, however, the performance is much higher, with recall values up to ~60% and precision values up to 100%. Further investigation revealed that the targets for which the prediction algorithm performed well (plasmepsin 1 and 2, histone deacetylase, DHFR and to a lesser extent, falcipain 2) were plasmodial orthologs of non-plasmodial protein targets. This finding is in agreement with previous studies, which have used orthologous proteins to extrapolate the prediction of bioactivities between target classes across species such as *P. falciparum* and *M. tuberculosis*.[47,48] However, these previous studies have not combined target prediction with PCM for MoA analysis, which is precisely the novelty of the approach presented here.

Target Prediction Performance for Plasmodial DHFR. The predictive performance of the target prediction algorithm was further investigated for the plasmodial target DHFR, where all 145 instances annotated on plasmodial DHFR were used as a test set. The top *n* predicted non-plasmodial targets were considered (*n* varied in the 1– 12 range), after which these targets were extrapolated to plasmodial targets (section "Domain-based extrapolation to *P. falciparum* targets" in Materials and Methods). For *n* in the 1–3 range, the recall values are 0%, 2.8% and 14,5%, respectively, whereas for *n* in the 4–7 range, the recall values are around 35%. The 2-fold CV resulted in a recall value of 79%. These results indicate that despite the fact that the training set did not contain any plasmodial bioactivity data, the model is still able to predict compounds active against plasmodial DHFR with 100% precision, based on bioactivity data for orthologous proteins across other species. The high precision value arises from the structural similarity of plasmodial DHFR

inhibitors and *T. gondii* DHFR inhibitors in the training set (the average MOLPRINT2D pairwise similarity between the *T. gondii* inhibitors and the plasmodial inhibitors was 16%, whereas the average pairwise similarity within the plasmodial dataset and the *T. gondii* dataset was 19% and 18% respectively). These results show the added benefit of incorporating domain-based extrapolation for target prediction purposes.

In addition, we found that varying the domain Tanimoto similarity cutoff between 0.5 and 1 did not alter the performance. Hence, in order to maintain high precision, a stringent domain Tanimoto similarity cutoff of 1 (i.e., requiring a 100% overlap in domain presence and absence between two proteins) was chosen and the top *n* predicted non-plasmodial targets considered was set to 4 for further analysis. Further investigation of the extrapolation from non-plasmodial targets to plasmodial targets revealed that only one protein class (*T. gondii* DHFR) was responsible for the extrapolation of predicted activities to plasmodial DHFR. As described earlier, there are clusters of *T. gondii* DHFR inhibitors that do not contain any plasmodial DHFR inhibitors (scaffolds identified in these clusters are shown in red boxes – **Figure 30** and clusters of plasmodial inhibitors that occupy space not covered by *T. gondii* inhibitors (5 scaffolds identified shown in green boxes in **Figure 30**)). Hence, for these clusters there is no overlap in scaffolds between both datasets. These observations explain the low recall of the model at this stage: plasmodial DHFR inhibitors located outside the space covered by *T. gondii* DHFR inhibitors are not retrieved by the model, thereby increasing the number of false negatives, whereas the plasmodial DHFR inhibitors that are present in the chemical space shared by inhibitors from both species are predicted with very high precision.

Adding plasmodial DHFR data to the training set drastically increased performance, more than doubling recall values to 79%, whereas precision values remained 100% (**Figure 31** – 2-fold CV).

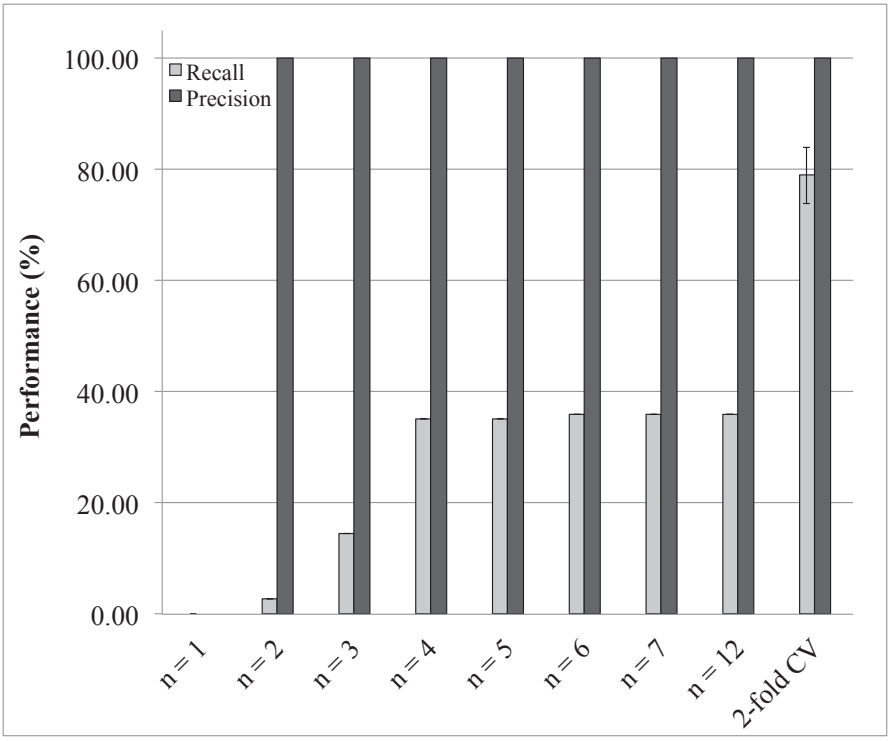


Figure 31. Performance of the DHFR target prediction model compared across a number of parameters. 145 data points annotated against plasmodial DHFR were used as a test set to assess the performance of the target prediction model. The top *n* predicted non-plasmodial targets were considered (*n* was varied for values between 1 and 12), after which these targets were extrapolated to plasmodial targets. When *n* increases, recall values rise up to 36% (with recall values of ~35% for $n = 3$ and $n = 4$). On the other hand, precision values are 100% for $n \geq 2$. The high precision values are likely to be explained by the fact that plasmodial DHFR inhibitors and *T. gondii* DHFR inhibitors occupy the same chemical space. In addition to varying the parameter *n*, we performed a 2-fold cross validation (averaged over 20 randomizations), which resulted in a drastic improvement as a recall value of 79% was achieved (with a standard deviation of 10.1%, which is shown as an error bar). These results show that domain-based extrapolations have added value to the prediction algorithm (correct predictions are made even when bioactivity data on plasmodial DHFR is not present in the training set) and that including plasmodial DHFR bioactivity data in the training set can drastically improve recall values.

Hence, this observation arises from the fact that the chemical space of the plasmodial DHFR inhibitors adds additional information corresponding to five new scaffolds (as highlighted in green boxes in **Figure 30**) to the model. However, despite the very high precision value achieved (100%), there is a drawback: given the great increase in recall value when novel scaffolds are

added to the dataset, the model is only able to correctly predict bioactivities for compounds with scaffolds that are already present in the training data. Hence, a diverse set of molecules is required in the training set in order to optimize recall values of the model. Given the benefit of both domain-based extrapolation and using plasmodial DHFR bioactivity data for model training, all plasmodial DHFR data were included in the training set for further MoA prediction of the GSK TCAMS phenotypic dataset in order to optimize recall values.

PCM Model Validation. The four algorithms used in this study (GBM, GP, RF and SVM) displayed similar performance on this dataset as the ranges of $RMSE_{test}$ and $R^2_{0\ test}$ differences are 0.04 pIC_{50} and 0.02 units, respectively. The GBM model exhibited the highest predictive ability with $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.79 and 0.59 pIC_{50} units respectively. Both internal and external validation metrics are given in **Table 6**.

Table 6. PCM, Family QSAR and Family QSAM performance on the PCM dataset. Abbreviations: QSAM Quantitative Structure-Activity Modeling, QSAR Quantitative Structure-Activity Relationship, GBM Gradient Boosting Machine, GP Gaussian Process, RF Random Forest, SVM Support Vector Machine. PCM, with $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.79 and 0.59 pIC_{50} units, outperforms both Family QSAR, with $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.63 and 0.78 pIC_{50} units, respectively, and Family QSAM, with with $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.09 and 1.22 pIC_{50} units, respectively.

	R^2_{CV}	$RMSE_{CV}$	$R^2_{0\ ext}$	$RMSE_{ext}$
GBM PCM	0.75	0.64	0.79	0.59
GP PCM	0.75	0.65	0.76	0.63
RF PCM	0.74	0.66	0.77	0.62
SVM PCM	0.76	0.63	0.77	0.62
Family QSAM	0.07	1.24	0.09	1.22
Family QSAR	0.61	0.80	0.63	0.78
Inductive Transfer	0.72	0.68	0.76	0.63

To ensure that the model's predictive ability was not the consequence of spurious correlations in the data, we trained ten GBM models with an increasingly higher percentage of the pIC_{50} values randomized. The performance of the ten models was assessed by examining the $RMSE_{test}$ and $R^2_{0\ test}$ values as a function of the level of randomization of the bioactivity values. The intercept was zero or negative when ~40% of the response variable was randomized. Therefore, the relationship established by the PCM

models between the descriptor space and the bioactivity values is not a consequence of chance correlations.[49]

PCM Outperforms One-space Models and IT on This Dataset. The Family QSAM model, trained on target descriptors only, displayed poor predictive ability with $RMSE_{test}$ and $R^2_{0\ test}$ values of 1.22 pIC_{50} units and 0.09, respectively (**Table 6**). By contrast, the Family QSAR model, trained on compound descriptors only, displayed satisfactory values for the statistical metrics according to our validation criteria, as the model exhibited $RMSE_{test}$ and $R^2_{0\ test}$ values of 0.78 pIC_{50} units and 0.63, respectively (**Table 6**). Hence, compound descriptors explain a large proportion of the variance, which may stem from the high correlation of the bioactivities of identical compounds against orthologs.

Furthermore, better performance is obtained for the GBM PCM model trained on amino acid descriptors and compound fingerprints, than for the GBM model trained on target identity fingerprints and compound fingerprints, with $RMSE_{test}$ values of 0.59 vs. 0.63 pIC_{50} units, respectively. This indicates that our selection of amino acid descriptors captured the binding site information of the different orthologs and thus allows explicit learning on this dataset (**Table 6**). Overall, these data suggest that the explicit inclusion of target information improves bioactivity prediction.

Several High-affinity DHFR Inhibitors Are Identified by Both Target Prediction and PCM. The targets for which the target prediction model had a class-specific F-measure higher than 40% were selected, leading to a shortlist of five proteins, namely: plasmepsin 1 and 2, histone deacetylase, DHFR and falcipain 2. Overall, a total of 1291 plasmodial predictions were made for 1017 compounds. DHFR is the most commonly predicted target, which represents 534 (41%) of the total predictions, followed by plasmepsin 1 (280 predictions – 22%) and plasmepsin 2 (273 predictions – 21 histone deacetylase (184 predictions – 14%) and falcipain 2 (20 predictions – 2%). Plasmodial DHFR has previously been proposed as a candidate target against resistant plasmodial strains.[50] In addition, the plasmepsin (1 and 2) and falcipain targets have previously been proposed as potential targets for anti-malarial therapy,[51] due to their involvement in the hemoglobin catabolism that occurs during the erythrocytic stage of the malarial parasite life cycle (plasmepsin proteins and falcipain proteins), and to their involvement in erythrocyte invasion and erythrocyte rupture (falcipain proteins).[52] Finally, plasmodial histone deacetylase has been proposed as a promising target for anti-malarial

therapy due to its key role in regulating gene transcription, and it has been shown that histone deacetylase inhibitors are potent inhibitors of the growth of *P. falciparum*.[53] Hence, there is sufficient evidence for all five predicted proteins for being a potential target.

In total, 534 compounds of the GSK TCAMS dataset were predicted to interact with DHFR, representing 3.95% of the total number of compounds in this dataset. Out of these 534 compounds, the predicted pIC_{50} values using PCM was 7 or greater for 25 compounds, between 6 and 7 for 92 compounds, and between 5 and 6 for 420. None of the 534 compounds was predicted to be inactive on DHFR (**Figure 32**).

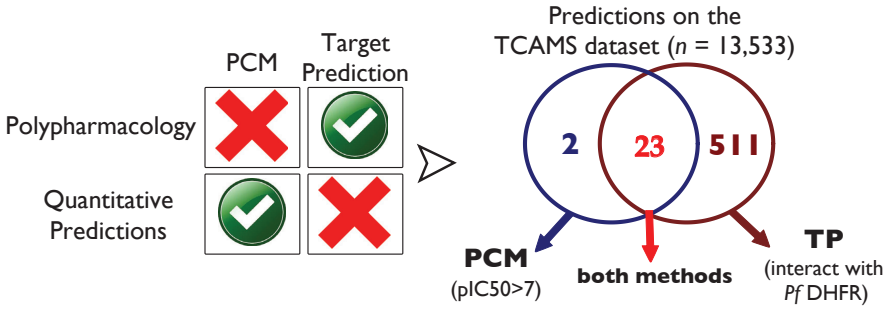


Figure 32. Complementarity between *in silico* target prediction and PCM. The target prediction algorithm predicted 534 compounds of the GSK TCAMS dataset to interact with DHFR, representing 3.95% of the total number of compounds in this dataset. Out of these 534 compounds, the PCM model predicted 23 compounds to have a pIC_{50} value of 7 or greater. Therefore, the combination of both methods permits the assessment of compound polypharmacology and provides quantitative bioactivity predictions.

Given that many of the compounds in ChEMBL are active in the low micromolar range, it is thus not surprising to obtain most of the predictions in this range.[54]

Interestingly, 23 of the 25 compounds with a predicted pIC_{50} value higher than 7 were already predicted to interact with DHFR by the target prediction algorithm (**Figure 32**) at the exclusion of any other target. The analysis of chemical scaffolds in the 25 compounds shows that only 2 scaffolds were identified, as 22 out of the 25 compounds (**Figure 33** – excluding compounds 137850, 123550 and 125380), share a common scaffold, namely: a 5-methylpyrido[2,3-d]pyrimidine-2,4-diamine ring with an aryl substituent in the 6-position.

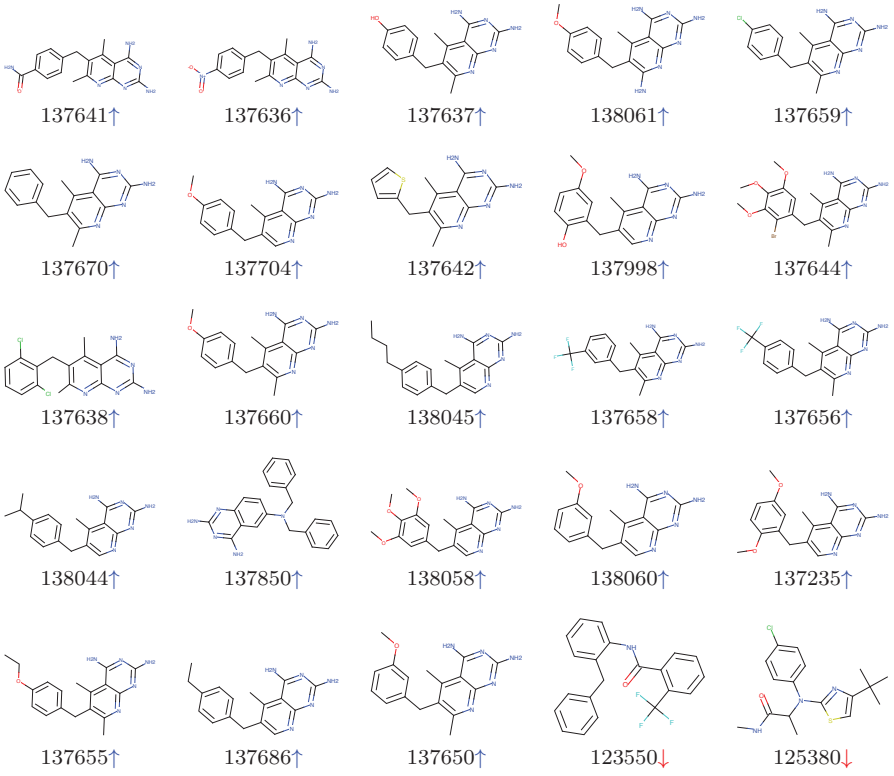


Figure 33. Compounds predicted to interact with DHFR by the target prediction algorithm, and predicted by the PCM model to have a pIC_{50} value higher than 7 pIC_{50} units. Compound IDs correspond to the TCMDC identifier given in the original dataset. The 23 compounds for which the IDs are accompanied by an upward-pointing arrow were identified by the two methods. The two compounds predicted to have a pIC_{50} value higher than 7 by the PCM model, but not predicted to interact with DHFR by the target prediction algorithm, are accompanied by a downward-pointing arrow. The 23 compounds predicted to be high-affinity DHFR inhibitors (upward-pointing arrows) share a common scaffold: a 5-methylpyrido [2,3-d]pyrimidine-2,4-diamine ring with an aryl substituent in the 6-position. Overall, it can be seen that these data indicate a high agreement between the target prediction algorithm and the PCM model to identify high-affinity DHFR inhibitors.

A methyl group or an amine group in the 7-position are also present in some compounds, such as 137637 and 138061, respectively. In all compounds with the common scaffold the aryl substituent is a phenyl ring with different substituents in the 3,4,5-positions, e.g., methoxy, hydroxy and carboxamide, except for compound 137642, which has 2-methyl-thiophene as aryl substituent.

Two additional compounds, 123550 and 125380 (**Figure 33**), predicted by PCM to display pIC_{50} values of 7.11 and 7.07, respectively, represent new

scaffolds. Remarkably, these two scaffolds were neither present in the PCM nor in the target prediction training set. Taken together, our results indicate a high agreement between the target prediction algorithm and the PCM model to identify high-affinity DHFR inhibitors. Using both methods simultaneously, it is possible to give higher priority to the compounds that are identified by both methods.

Conclusions

In this study, the complementarity of *in silico* target predictions and proteochemometric modelling (PCM) was evaluated for the retrospective identification of *P. falciparum* DHFR inhibitors. The target prediction algorithm exhibited respective recall and precision values of 79% and 100% for plasmodial DHFR. The high precision value is explained by the structural similarity of plasmodial and the *T. gondii* DHFR inhibitors, which were part of the training set and were found to be relevant for extrapolation (the average MOLPRINT2D pairwise similarity between the *T. gondii* inhibitors and the plasmodial inhibitors was 16%, whereas the average pairwise similarity within the plasmodial dataset and the *T. gondii* dataset was 19% and 18% respectively).

We showed that high-affinity inhibitors from the GSK TCAMS phenotypic dataset are independently identified by both methods: 534 compounds from the GSK TCAMS dataset were identified as DHFR inhibitors by the target prediction algorithm, whereas the PCM algorithm identified 25 high affinity compounds, 23 of which were already identified by the target prediction algorithm. The combination of both methods permits the assessment of compound polypharmacology and provides insight into the potency/affinity of small molecules.

We presented an approach that can be potentially extended to other human, bacterial or plasmodial targets. The inherent capability of PCM to combine bioactivity data for related targets, even for targets spanning distant phyla, is likely to improve the mining of currently available multi-target bioactivity databases. Similarly, domain-based extrapolation permits *in silico* target predictions to be extended to non-mammalian orthologous proteins for which less bioactivity data is usually available.

References

(1) Jalencas, X., and Mestres, J. (2013) On the origins of drug polypharmacology. *Med. Chem. Commun. 4*, 80–87.

(2) Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Cote, S., Shoichet, B. K., and Urban, L. (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature 486*, 361–367.

(3) Cortés-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Mendez-Lucio, O., IJzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T., Van Westen, G. J. P., and Bender, A. (2015) Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *Med. Chem. Commun. 6*, 24-50.

(4) Van Westen, G. J., Swier, R. F., Cortés-Ciriano, I., Wegner, J. K., Overington, J. P., IJzerman, A. P., Van Vlijmen, H. W., and Bender, A. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *J. Cheminform. 5*, 42-61.

(5) Poroikov, V., Filimonov, D., Lagunin, A., Gloriozova, T., and Zakharov, A. (2007) PASS: identification of probable targets and mechanisms of toxicity. *SAR QSAR Environ. Res. 18*, 101–110.

(6) Nidhi, Glick, M., Davies, J. W., and Jenkins, J. L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model. 46*, 1124–1133.

(7) Nigsch, F., Bender, A., Jenkins, J. L., and Mitchell, J. B. O. (2008) Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model. 48*, 2313–2325.

(8) Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol. 25*, 197–206.

(9) Wale, N., and Karypis, G. (2010) Target Fishing for Chemical Compounds using Target-Ligand Activity data and Ranking based Methods. *J. Chem. Inf. Model. 49*, 2190–2201.

(10) 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.

(11) 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.

(12) Perlmann, P., and Troye-Blomberg, M. (2000) Malaria blood-stage infection and its conyrol by the immune system. *Folia Biol. 46*, 210–218.

(13) Olliaro, P. (2001) Mode of action and mechanisms of resistance for antimalarial drugs. *Pharmacol. Ther. 89*, 207–219.

(14) Hecht, D., and Fogel, G. B. (2012) Modeling the evolution of drug resistance in malaria. *J. Comput. Aided Mol. Des. 26*, 1343–1353.

(15) Moran, M., Guzman, J., and Ropars, A.-L. (2007) The malaria product pipeline: planning for the future. *Georg. Inst. Int. Heal.*

(16) ChEMBL - Neglected Tropical Disease.

(17) 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.

(18) Verma, R., Tiwari, A., Kaur, S., Varshney, G. C., and Raghava, G. P. S. (2008) Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *BMC Bioinformatics 9*, 201–211.

(19) Jamal, S., Periwal, V., and Scaria, V. (2013) Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. *BMC Bioinformatics 14*, 2105–2114.
(20) Subramaniam, S., Mehrotra, M., and Gupta, D. (2011) Support Vector Machine Based Prediction of P. falciparum Proteasome Inhibitors and Development of Focused Library by Molecular Docking. *Comb. Chem. High Throughput Screen. 14*, 898–907.
(21) *Dotmatics Vortex*, version 2013.03.20719, Dotmatics: The Old Monastery, Windhill, Bishops Stortford, Herts, U.K., 2013.
(22) Bender, A. (2010) Databases: Compound bioactivities go public. *Nat. Chem. Biol. 6*, 309–309.
(23) 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.
(24) *ChemAxon Standardizer*, version 5.12; ChemAxon, Ltd: Budapest, Hungary, 2012.
(25) Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. a, Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C., and Yong, S.-Y. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucl. Acids Res. 40*, D306–D312.
(26) The Uniprot Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucl. Acids Res. 41*, D43–D47.
(27) Bender, A., Mussa, H. Y., and Glen, R. C. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Model. 44*, 170–178.
(28) Bender, A., Mussa, H. Y., and Glen, R. C. (2004) Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Model. 44*, 1708–1718.
(29) Sastry, M., Lowrie, J. F., Dixon, S. L., and Sherman, W. (2010) Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model. 50*, 771–784.
(30) O'Boyle, N. M., Morley, C., and Hutchison, G. R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J. 2*, 5–11.
(31) Crisman, T. J., Parker, C. N., Jenkins, J. L., Scheiber, J., Thoma, M., Kang, Z. Bin, Bender, A., Nettles, J. H., Davies, J. W., and Glick, M. (2007) Understanding false positives in reporter gene assays: in silico chemogenimics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model. 47*, 1319–1327.
(32) Bender, A., Mikhailov, D., Glick, M., Scheiber, J., Davies, J. W., Cleaver, S., Marshall, S., Tallarico, J., Harrington, E., Cornella-Taracido, I., and Jenkins, J. (2009) Use of Ligand Based Models for Protein Domains To Predict Novel Molecular Targets and Applications To Triage Affinity Chromatography Data. *J. Proteome Res. 8*, 2575–2585.
(33) Prathipati, P., Ma, N., Manjunatha, U. H., and Bender, A. (2009) Fishing the target of antitubercular compounds: in silico target deconvolution model development and validation. *J. Proteome Res. 8*, 2788–2798.
(34) Murrell, D. S., Cortés-Ciriano, I., Van Westen, G. J. P., Malliavin, T., and Bender, A. (2014) Chemistry Aware Model Builder (camb): an R Package for Predictive Bioactivity Modeling.
(35) Cody, V., Galitsky, N., Luft, J. R., Pangborn, W., Rosowsky, A., and Blakley, R. L. (1997) Comparison of two independent crystal structures of human dihydrofolate reductase ternary

complexes reduced with nicotinamide adenine dinucleotide phosphate and the very tight-binding inhibitor PT523. *Biochemistry 36*, 13897–13903.
(36) Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol. 7*, 539–544.
(37) Consonni, V., Ballabio, D., and Todeschini, R. (2010) Evaluation of model predictive ability by external validation techniques. *J. Chemom. 24*, 194–201.
(38) Hawkins, D. M., Basak, S. C., and Mills, D. (2003) Assessing Model Fit by Cross-Validation. *J. Chem. Inform. Comput. Sci. 43*, 579–586.
(39) Tropsha, A., Gramatica, P., and Gombar, V. (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci. 22*, 69–77.
(40) Golbraikh, A., and Tropsha, A. (2002) Beware of q2! *J. Mol. Graph. Model. 20*, 269–276.
(41) Brown, J. B., Okuno, Y., Marcou, G., Varnek, A., and Horvath, D. (2014) Computational chemogenomics: Is it more than inductive transfer? *J. Comput. Aided Mol. Des. 28*, 597–618.
(42) Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol. 4*, e1000173.
(43) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat. 29*, 1189–1232.
(44) Rasmussen, C. E., and Williams, C. K. I. (2006) Gaussian Processes for Machine Learning. The MIT Press.
(45) Breiman, L. (2001) Random Forests. *Mach. Learn. 45*, 5–32.
(46) Kuhn, M. (2008) Building Predictive Models in R Using the caret Package. *J. Stat. Softw. 28*, 1–26.
(47) Spitzmüller, A., and Mestres, J. (2013) Prediction of the P. falciparum target space relevant to malaria drug discovery. *PLoS Comput. Biol. 9*, e1003257.
(48) Martínez-Jiménez, F., Papadatos, G., Yang, L., Wallace, I. M., Kumar, V., Pieper, U., Sali, A., Brown, J. R., Overington, J. P., and Marti-Renom, M. A. (2013) Target prediction for an open access set of compounds active against Mycobacterium tuberculosis. *PLoS Comput. Biol. 9*, e1003253.
(49) Clark, R. D., and Fox, P. C. (2004) Statistical variation in progressive scrambling. *J. Comput. Aided Mol. Des. 18*, 563–576.
(50) Yuthavong, Y., Tarnchompoo, B., Vilaivan, T., Chitnumsub, P., Kamchonwongpaisan, S., Charman, S. A., McLennan, D. N., White, K. L., Vivas, L., Bongard, E., Thongphanchang, C., Taweechai, S., Vanichtanankul, J., Rattanajak, R., Arwon, U., Fantauzzi, P., Yuvaniyama, J., Charman, W. N., and Matthews, D. (2012) Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. *Proc. Natl. Acad. Sci. U.S.A. 109*, 16823–16828.
(51) Ersmark, K., Samuelsson, B., and Hallberg, A. (2006) Plasmepsins as Potential Targets for New Antimalarial Therapy. *Med. Res. Rev. 26*, 626–666.
(52) Marco, M., and Coterón, J. M. (2012) Falcipain inhibition as a promising antimalarial target. *Curr. Top Med. Chem. 12*, 408–444.
(53) Andrews, K. T., Tran, T. N., Wheatley, N. C., and Fairlie, D. P. (2009) Targeting histone deacetylase inhibitors for anti-malarial therapy. *Curr. Top Med. Chem. 9*, 292–308.
(54) Cortés-Ciriano, I., Koutsoukas, A., Abian, O., Glen, R. C., Velazquez-Campoy, A., and Bender, A. (2012) Experimental validation of in silico target predictions on synergistic protein targets. *Med. Chem. Commun. 4*, 278–288.

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.**

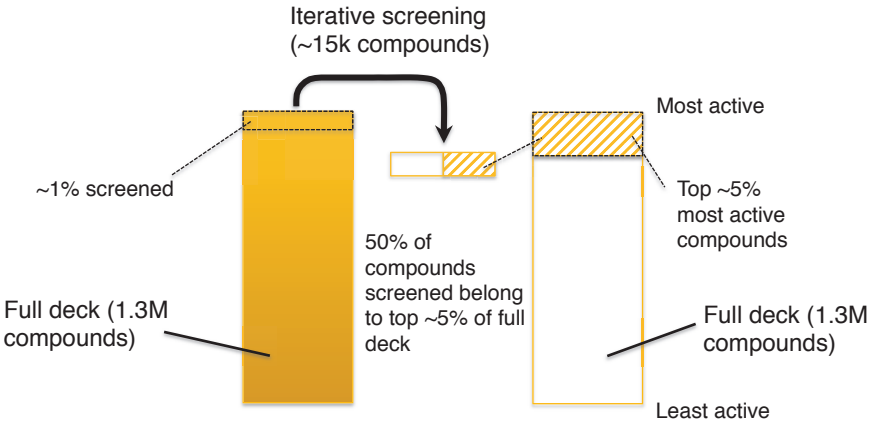


Figure 34. Illustration of efficiency gains using iterative screening. Half of the compounds selected iteratively, which corresponds to only 0.5% of the entire screening collection of 1.3 million compounds (full deck), were found to be among the top 5% of compounds in terms of activity. This indicates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

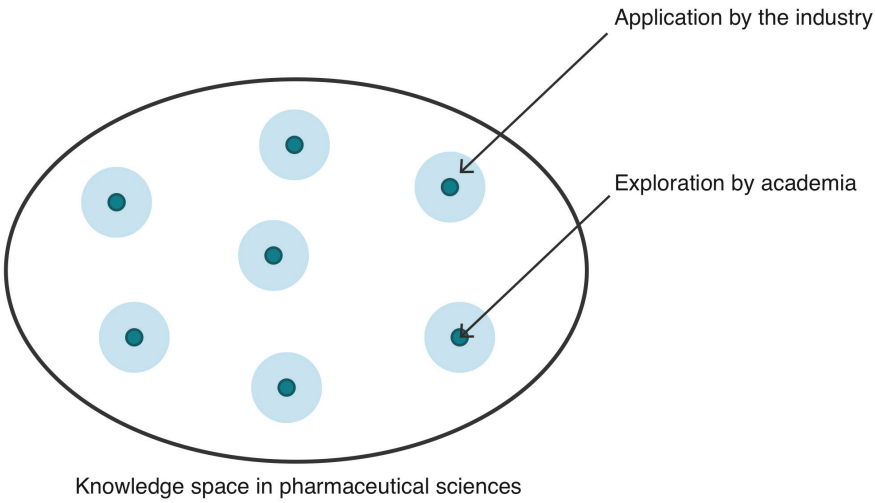


Figure 35. Exploration of knowledge space in pharmaceutical sciences by collaborative efforts of academia and the industry. Academia could focus on research areas that are not necessarily profit-driven (e.g., neglected diseases and drug repositioning) and therefore more likely to be neglected by the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

Summary

This thesis describes various analyses of life science data with the aim of achieving efficiency gains in future experimental campaigns and novel insights into compound mode-of-action (i.e., the protein target modulated for the desired phenotypic effect). The increase in publicly available life science data has created opportunities for bioactivity modeling, and the role cheminformatics and bioinformatics play in the latter is discussed.

Chapter one describes the relevance of computational drug discovery. The fundamentals of bioactivity modeling are explained in detail, followed by an introduction of the topics discussed in this thesis. *Chapter two* is a literature review on data-driven approaches used for compound library design, hit triage and bioactivity modeling in high-throughput screening. High-throughput screening campaigns are routinely performed in pharmaceutical companies to explore activity profiles of chemical libraries for the identification of promising candidates for further investigation. In particular, the remarkable progress in the activity modeling area since the recent introduction of large-scale bioactivity-based compound similarity metrics is discussed in detail, outlining its significance in the field.

In *Chapter three*, the relevance of chemical space for bioactivity modeling is inspected. Chemicals can be described in terms of a set of characteristics (descriptor) that computers can easily use to assess similarity between molecules. Chemical diversity is a widely applied concept used to select structurally diverse subsets of molecules, often with the objective of maximizing the number of hits in biological screening. The extent to which the descriptors used in this study correlated in their assessment of molecular diversity across a number of compound sets ranging in size, diversity and origin is outlined in detail. Descriptors based on atom topology are shown to correlate well in rank-ordering compounds, whereas shape-based descriptors show weak correlation with other descriptor types. Finally, the descriptor "Bayes Affinity Fingerprints" which is based on predicted bioactivity profiles of compounds is shown to be most effective in selecting compound sets that are diverse in bioactivity space.

Chapter four illustrates the application of a computational method geared toward systematic compound prioritization, aimed at increasing the efficiency of compound screening campaigns over high-throughput screening campaigns performed currently in the pharmaceutical industry. The screening strategy described in this chapter consisted of the iterative selection of compounds chemically and biologically similar to actives identified in

multiple rounds of testing and was retrospectively validated on Novartis high-throughput screening data. Large efficiency gains were observed across assays covering a wide range of assay biology: by only screening 1% of the full screening collection, a consistent retrieval of diverse sets of compounds belonging to the top 0.5% was achieved. Employing this method can potentially lead to considerable savings in both time and resources.

Chapter five describes the data-driven derivation of an "informer compound set". Once screened, this set provides the most information on which yet untested compounds from the remainder of a large compound collection to screen next, irrespective of biological target. The derivation of this informer set involves the concept of *active learning*, which attempts to maximize the predictive power of the informer set. A retrospective validation of this set was performed on public high-throughput screening data, and an improvement in early retrieval of active compounds is observed for 38 out of 46 assays, increasing the success rate of smaller follow-up screens.

The final research chapter, *Chapter six*, represents a case study in the context of mode-of-action analysis of anti-malarial compounds identified in phenotypic screens by GlaxoSmithKline. Here, the application of two machine learning methods (Bayesian target prediction and proteochemometrics modeling) is illustrated for simultaneous polypharmacology and affinity predictions. Overall, 534 compounds were identified as dihydrofolate reductase inhibitors by the target prediction algorithm, while the proteochemometrics modeling approach identified 25, with an overlap of 23 compounds between both methods.

Finally, *Chapter seven* draws general conclusions from this thesis and provides future perspectives where some of my views on (early) drug discovery in academia and the pharmaceutical industry are discussed.

Samenvatting

Dit proefschrift beschrijft verschillende analyses van data uit de disciplines chemie, biologie en biochemie met het doel om de efficiëntie van experimentele testen voor het ontdekken van nieuwe geneesmiddelen te verbeteren en om nieuwe inzichten te verkrijgen in het mechanisme van bioactiviteit van moleculen. De recente toename in de hoeveelheid publieke data heeft geleid tot nieuwe kansen op het gebied van bioactiviteitmodellering, en de rol die de cheminformatics en de bioinformatics hierin spelen wordt toegelicht.

Hoofdstuk één beschrijft het belang van de computational drug discovery. De principes van bioactiviteitmodellering worden in detail uitgelegd, gevolgd door een inleiding tot de onderwerpen die in dit proefschrift aan bod komen.

Hoofdstuk twee is een recensie over data-gedreven methodes die gebruikt worden bij het ontwerpen van molecuulsets voor experimentele testen, hit triage en bioactiviteitmodellering in high-throughput screening. High-throughput screening is een techniek die regelmatig wordt toegepast in de farmaceutische industrie om de activiteitsprofielen van molecuulsets te verkennen, met als doel de identificatie van mogelijk actieve stoffen voor verder onderzoek. De opmerkelijke vooruitgang op het gebied van bioactiviteitmodellering sinds de recente introductie van grootschalige moleculaire similarity metrics gebaseerd op bioactiviteitsprofielen wordt in het bijzonder besproken.

In *Hoofdstuk drie* wordt het belang van de chemische ruimte voor bioactiviteitmodellering onderzocht. Moleculen kunnen beschreven worden als een reeks eigenschappen (descriptor) die computers kunnen gebruiken om de gelijkenis tussen moleculen vast te stellen. Het concept van de chemische ruimte wordt vaak toegepast bij het selecteren van structureel diverse moleculen, met het doel het aantal actieve stoffen in biologische experimenten te optimaliseren. De mate waarin de descriptoren die in dit onderzoek gebruikt werden correleerden bij het vaststellen van moleculaire diversiteit wordt uitgebreid uitgelegd. Descriptoren die gebaseerd zijn op atoomtopologie vertonen grote overeenkomsten in hun beoordeling van moleculaire diversiteit, terwijl de op vorm gebaseerde descriptoren weinig overeenkomsten vertoonden met andere descriptortypen. Tenslotte werd gevonden dat de descriptor “Bayes Affinity Fingerprints”, die gebaseerd is op voorspelde bioactiviteitsprofielen van moleculen, het meest effectief is voor het selecteren van molecuulsets die divers zijn in de bioactiviteitsruimte.

In *Hoofdstuk vier* is de toepassing van computertechnieken om op systematische wijze moleculen te prioriteren beschreven, met als doel de efficiëntie van experimentele testen te verbeteren ten opzichte van high-throughput screening. De methode die in dit hoofdstuk beschreven wordt bestond uit de iteratieve selectie van moleculen die chemische en biologische gelijkenis vertoonden met de in eerdere iteraties geïdentificeerde actieve moleculen. De methode werd op retrospectieve wijze gevalideerd op Novartis high-throughput screening data en leidde tot een grote verbetering van efficiëntie in diverse assays: door slechts 1% van het aantal stoffen te testen werden consequent diverse molecuulsets teruggevonden die behoorden tot de top 0.5% qua activiteit. Het gebruik van deze methode kan mogelijk leiden tot aanzienlijke besparingen in tijd en middelen.
In *Hoofdstuk vijf* wordt de constructie van een "informer compound set" met behulp van data-gedreven methodes besproken. Deze informer set verschaft de meeste informatie over welke ongeteste stoffen uit een grote molecuulset het beste getest kunnen worden in een volgende ronde, ongeacht het biologische target. Het concept van *active learning* werd gebruikt voor de afleiding van de informer set, waardoor de hoeveelheid informatie van de set wordt vergroot. Een retrospectieve validatie van de informer set werd uitgevoerd op publieke high-throughput screening data, en een verbetering in vroege herkenning van actieve stoffen werd bereikt in 38 van de 46 assays.
Hoofdstuk zes beschrijft een case study over de analyse van het mechanisme van bioactiviteit van mogelijk actieve stoffen tegen malaria, welke ontdekt werden in fenotypische testen bij GlaxoSmithKline. De toepassing van twee machine learning methodes (Bayesian target prediction en proteochemometrics modeling) voor de gelijktijdige voorspelling van polyfarmacologie en affiniteit wordt beschreven. Het target prediction algoritme identificeerde 534 dihydrofolate reductase inhibitoren, terwijl de proteochemometrics modeling techniek 25 inhibitoren identificeerde, met een overlap van 23 moleculen tussen beide methoden.
Tenslotte worden er in *Hoofdstuk zeven* algemene conclusies getrokken uit het onderzoek dat in dit proefschrift beschreven wordt gevolgd door mijn toekomstperspectief over het vroege stadium van geneesmiddelenonderzoek in de academie en de farmaceutische industrie.

List of Publications

Paricharak, S., Méndez-Lucio, O., Ravindranath, A. C., Bender, A., IJzerman, A. P., and Van Westen, G. J. P. Data-driven approaches used for compound library design, hit triage, and bioactivity modeling in high-throughput screening. *Submitted*

Paricharak, S., IJzerman, A. P., Jenkins, J. L., Bender, A., and Nigsch, F. (2016) Data-driven derivation of an "informer compound set" for improved selection of active compounds in high-throughput screening. *J. Chem. Inf. Model. DOI: 10.1021/acs.jcim.6b00244*.

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.

Mohan, C. D., Srinivasa, V., Rangappa, S., Mervin, L., Mohan, S., Paricharak, S., Baday, S., Li, F., Shanmugam, M. K., Chinnathambi, A., Zayed, M. E., Alharbi, S. A., Bender, A., Sethi, G., Basappa, and Rangappa, K. S. (2016) Trisubstituted-imidazoles induce apoptosis in human breast cancer cells by targeting the oncogenic PI3K/Akt/mTOR signaling pathway. *PLoS One 11*, e0153155.

Kumar, K. H., Paricharak, S., Mohan, C. D., Bharathkumar, H., Nagabhushana, G. P., Rajashekar, D. K., Chandrappa, G. T., Bender, A., Basappa, and Rangappa, K. S. (2016) Nano-MoO3-mediated synthesis of bioactive thiazolidin-4-ones acting as anti-bacterial agents and their mode-of-action analysis using in silico target prediction, docking and similarity searching. *New J. Chem. 40*, 2189–2199.

Anusha, S., CP, B., Mohan, C. D., Mathai, J., Rangappa, S., Mohan, S., Chandra, Paricharak, S., Mervin, L., Fuchs, J. E., M, M., Bender, A., Basappa, and Rangappa, K. S. (2015) A nano-MgO and ionic liquid-catalyzed "green" synthesis protocol for the development of adamantyl-imidazolo-thiadiazoles as anti-tuberculosis agents targeting sterol 14α-demethylase (CYP51). *PLoS One 10*, e0139798.

Paricharak, S., Cortés-Ciriano, I., IJzerman, A. P., Malliavin, T. E., and Bender, A. (2015) Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity of small molecules. *J. Cheminform. 7*, 15–25.

Neelgundmath, M., Dinesh, K. R., Mohan, C. D., Li, F., Dai, X., Siveen, K. S., Paricharak, S., Mason, D. J., Fuchs, J. E., Sethi, G., Bender, A., Rangappa, K. S., Kotresh, O., and Basappa. (2015) Novel synthetic coumarins that target NF-κB in hepatocellular carcinoma. *Bioorg. Med. Chem. Lett. 25*, 893–897.

Bharathkumar, H., Paricharak, S., Dinesh, K. R., Siveen, K. S., Fuchs, J. E., Rangappa, S., Mohan, C. D., Mohandas, N., Kumar, A. P., Sethi, G., Bender, A., Basappa, and Rangappa, K. S. (2014) Synthesis, biological evaluation and in silico and in vitro mode-of-action analysis of novel dihydropyrimidones targeting PPAR-γ. *R. Soc. Chem. Adv. 4*, 45143–45146.

Bharathkumar, H., Sundaram, M. S., Jagadish, S., Paricharak, S., Hemshekhar, M., Mason, D., Kemparaju, K., Girish, K. S., Basappa, Bender, A., and Rangappa, K. S. (2014) Novel benzoxazine-based aglycones block glucose uptake in vivo by inhibiting glycosidases. *PLoS One 9*, e102759.

Koutsoukas, A., Paricharak, S., Galloway, W. R. J. D., Spring, D. R., IJzerman, A. P., Glen, R. C., Marcus, D., and Bender, A. (2014) How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model. 54*, 230–242.

Paricharak, S., Klenka, T., Augustin, M., Patel, U. A., and Bender, A. (2013) Are phylogenetic trees suitable for chemogenomics analyses of bioactivity data sets: the importance of shared active compounds and choosing a suitable data embedding method, as exemplified on kinases. *J. Cheminform. 5*, 49–68.

Curriculum Vitae

Shardul Paricharak was born in Willemstad, Curaçao, on 21st of July 1989. He went to high school at Radulphus College and graduated as the best VWO (highest level) student of Curaçao in 2007. He then moved to the Netherlands to pursue his undergraduate study Life Science & Technology at TU Delft and Leiden University, which he graduated *cum laude*. During this study, he spent one semester at Lund University (Sweden) to study bioinformatics and passed all courses with distinction. This exchange program sparked his interests in the computational side of biology and led him to pursue his undergraduate research internship at the Division of Medical Pharmacology at Leiden University, where he analyzed glucocorticoid signaling pathways in the brain using bioinformatics tools. He then gained some industrial experience during a summer job at Galapagos B.V. (The Netherlands), where he worked on developing proprietary software to design siRNA. Following this, he performed two graduate research internships in bioinformatics and cheminformatics at Utrecht University (The Netherlands) and the University of Cambridge (UK), respectively, where he worked on personalized genomics data analysis and kinase bioactivity modeling. He received the Fundatie van Renswoude grant and the Huygens Scholarship for academic excellence for his work at the University of Cambridge.

In 2012, Shardul obtained his master's degree in pharmaceutical sciences from Utrecht University, with a GPA score of 4.0. In the same year, he obtained an NWO Mosaic personal grant for a PhD at Leiden University and the University of Cambridge, during which he worked on cheminformatics data analyses and machine learning. He obtained Dr Hendrik Muller's Vaderlandsch Fonds scholarship and the Prins Bernhard Cultuurfonds Pieter Beijer scholarship in support of his PhD and fellowship at King's College (Cambridge) as an affiliate member, respectively. He gained some further industrial experience at Novartis in Switzerland, where he worked on applying computational methods to improve the efficiency of high-throughput screens. Shardul has presented his research at the 13th LCDS meeting: combining data science and drug discovery (Leiden, The Netherlands), the Big data in medicine conference (Cambridge, UK), the Cambridge Cheminformatics Meeting (Cambridge, UK), the Scandinavian Symposium on Chemometrics (Chia, Italy), the Towards New Therapeutics for Diseases of the Developing World conference, (Tres Cantos, Spain), and the 8th German Conference on Chemoinformatics (Goslar, Germany).

Acknowledgements

I feel incredibly fortunate to have had the opportunity to conduct research at three institutions across three countries during my PhD, collaborating with world-class scientists in both academia and the pharmaceutical industry. I would like to thank the Netherlands Organisation for Scientific Research (NWO), the Prins Bernhard cultuurfonds and Novartis for financial support. My supervisors Ad IJzerman, Andreas Bender and Florian Nigsch are thanked for invaluable scientific contributions, helpful advice and mentoring during my PhD.

My colleagues at the University of Cambridge, Leiden University and Novartis are thanked for insightful discussions, pleasant shared experiences, continuous support and friendship. In particular, I would like to thank Aakash Chavan Ravindranath, Oscar Méndez-Lucio, Richard Lewis, Alexios Koutsoukas, Basappa, Isidro Cortés-Ciriano and Gerard van Westen for fruitful collaborations. My students Richard Lewis, Oana Diaconescu and Chi Chung Lam are thanked for providing me the opportunity to develop my leadership skills.

From the personal side, I would like to thank my parents, Seema Paricharak and Atul Paricharak, and my brother Nakul Paricharak for unconditional support throughout my PhD, during both ups and downs. Finally, I would also like to thank my parents for striving hard to provide me the opportunity of pursuing academic education, which I value immensely and believe will tremendously benefit me in my future endeavors.