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

Paricharak, S.A.

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**.

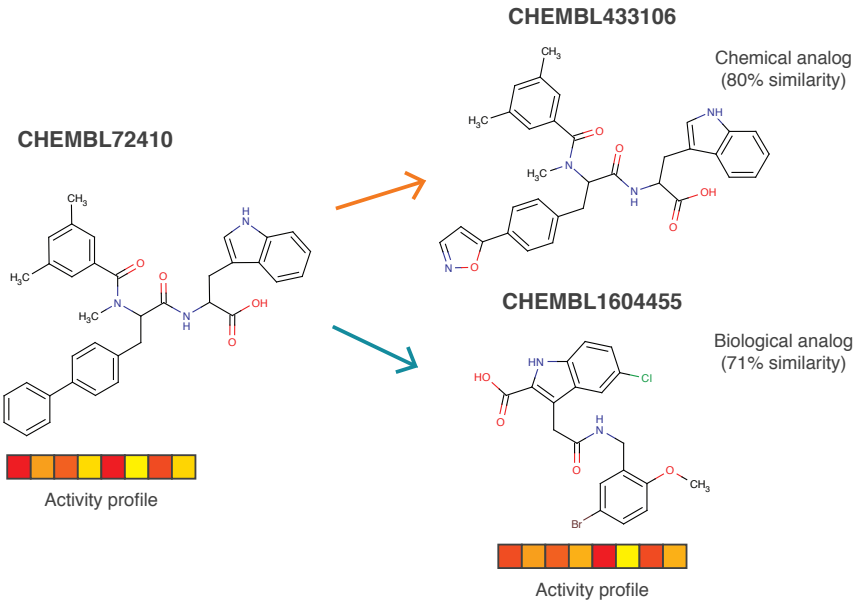

Figure 1. An example of a similarity search originating from research conducted in *Chapter four*, based on molecular structure (chemical descriptor – above) and on activity profile (biological descriptor – below). Chemical analogs often have the same or a very similar scaffold[17] (i.e., molecular framework), 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.

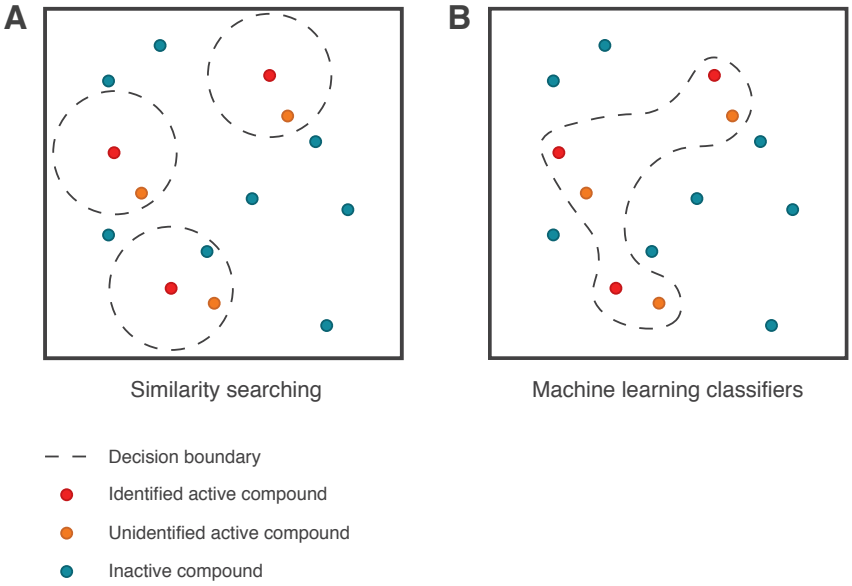

Figure 2. Hypothetical decision boundaries constructed by similarity searching (A) versus those constructed using classification methods (B). Compounds present in the area within the decision boundaries are classified as active according to the method used. Similarity searching assumes that instances 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]

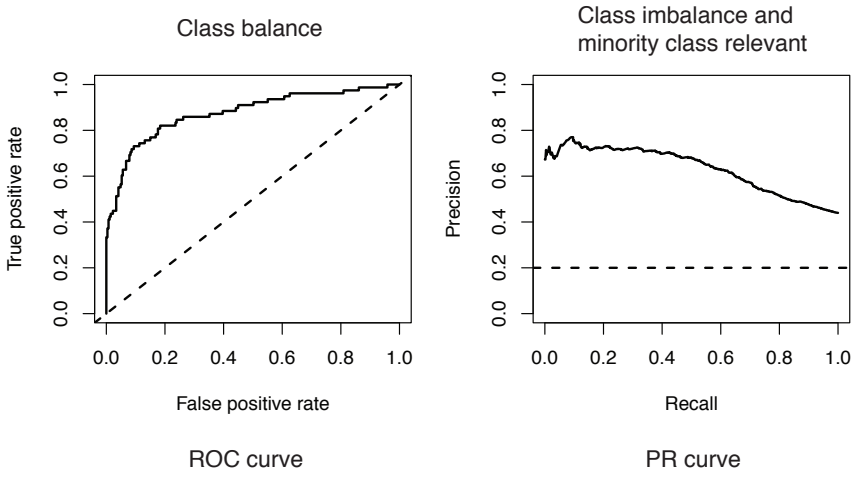

Figure 3. The receiver operating characteristic (ROC) curve (left) and the precision-recall (PR) curve (right). The area under these curves (AUC) is a metric for classifier performance across a range of cutoffs, resulting in a number between zero and one. The dashed lines in both plots represent random 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**).

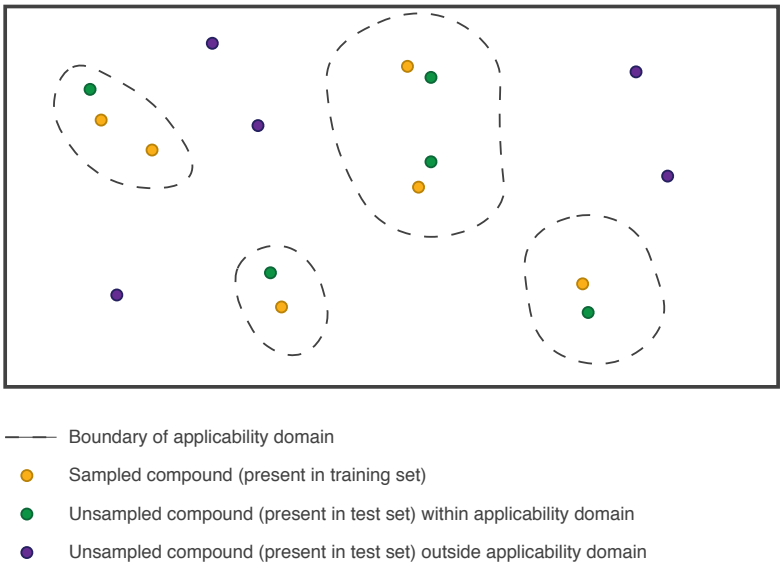

Figure 4. Applicability domain of a classification model. Sampled compounds (yellow) allow the classifier to understand the neighboring descriptor space, and test compounds in that space (green) fall within the applicability domain and can in theory be predicted with more confidence. Areas that are undersampled 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.