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

Paricharak, S.A.

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]

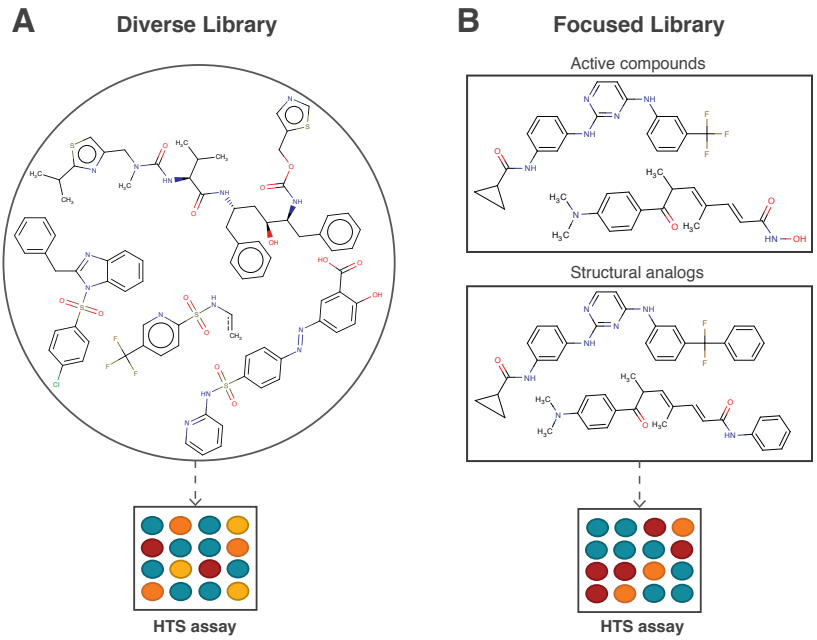

Figure 5. Diverse libraries compared to focused libraries. Structurally diverse libraries are used to efficiently explore relevant chemical space for targets with few known active chemotypes or for phenotypic assays (A).[35] This is performed to provide multiple starting points for further development. 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]

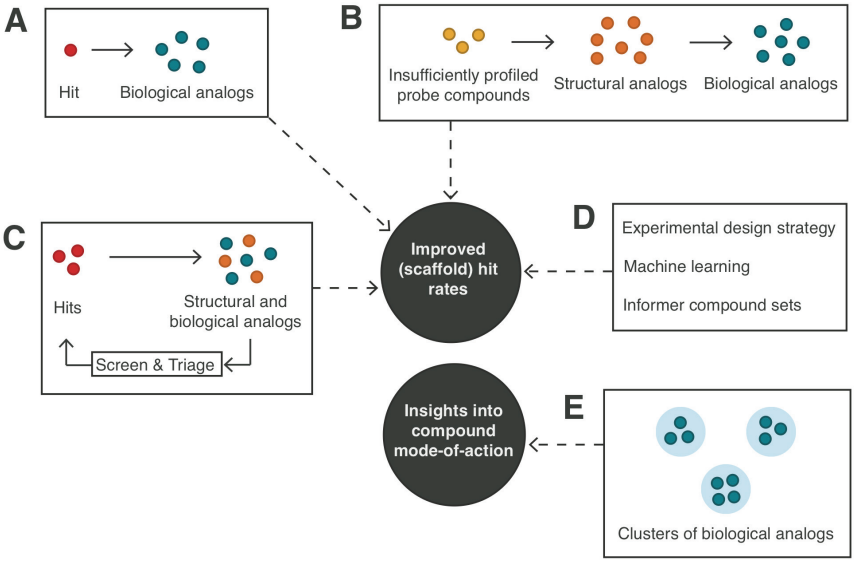

Figure 6. Overview of recent studies improving (scaffold) hit rates and providing insights into compound mode-of-action. Describing compound bioactivity across ~200 assays at Novartis, Petrone *et al.*[25] took the concept of bioactivity-based similarity to an unparalleled level. Here, biological analogs 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.