JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

pubs.acs.org/jcim

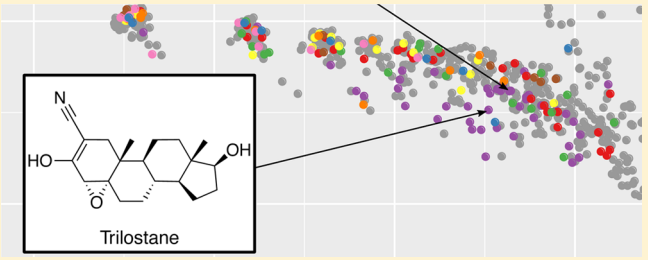
bioassayR: Cross-Target Analysis of Small Molecule Bioactivity

Tyler William H. Backman and Thomas Girke*

Institute for Integrative Genome Biology, University of California, Riverside, Riverside, California 92521, United States

Supporting Information

ABSTRACT: Despite a large and rapidly growing body of small molecule bioactivity screens available in the public domain, systematic leverage of the data to assess target druggability and compound selectivity has been confounded by a lack of suitable cross-target analysis software. We have developed bioassayR, a computational tool that enables simultaneous analysis of thousands of bioassay experiments performed over a diverse set of compounds and biological targets. Unique features include support for large-scale cross-target analyses of both public and custom bioassays, generation of high throughput screening fingerprints (HTSFPs), and an optional preloaded database that provides access to a substantial portion of publicly available bioactivity data. bioassayR is implemented as an open-source R/Bioconductor package available from https://bioconductor.org/packages/bioassayR/.



INTRODUCTION

Diverse collections of small molecules have been screened over the past decade against a wide array of distinct protein target families. The resulting high throughout screening (HTS) data are available in community driven databases such as PubChem Bioassay, ChEMBL, ZINC, ChemDB, and many others (list in Table S1 of the Supporting Information).[1−4] As demonstrated by many data mining efforts, these bioactivity resources provide an opportunity for studying the selectivity patterns and molecular mechanisms of small molecule−target interactions on a broad scale.[1,5−11] These insights have the potential to lead to the discovery of drug candidates and protein target sites relevant for medical or chemical genomics applications. The data can also be used to identify and exclude drug candidates with largely unselective binding properties (e.g., promiscuous binders) that have been found to be of limited use to most application areas.[8,12,13] Moreover, the bioactivity data can be used to develop multitarget treatments specific to one or several cross-connected pathways; to identify alternative uses for existing drugs; or to predict potential side and toxic effects.[14−16] Data from single target screens (i.e., a bioassay with a specific target protein) can also be helpful for prioritizing potential target sites in multiplexed or high-content screens, where a specific target protein is usually unknown. Furthermore, large-scale compound bioassay data can be used to create an inventory of molecular functions and proteins that are accessible or resistant to perturbations by small molecules. These "druggability profiles" can be used to guide decision processes in selecting the most efficient target sites for a specific research application in drug discovery and other small molecule driven research disciplines.[17]

Most of the small molecule bioactivity data available in the above-mentioned public databases were generated by systematic screening efforts of the Molecular Libraries Program (MLP), the Chemical Biology Program of the Broad Institute, and a variety of smaller public efforts.[18] The online interfaces of these databases provide many useful search and download options for focused analysis of a small number of molecules or target proteins.[1,2] Although several projects have developed statistical methods and sample scripts applicable to cross-target analysis, there is currently no general purpose software infrastructure available to perform these tasks in a systematic and fully customizable manner.[5−9,13,19]

To address this deficit, we have developed bioassayR, a computational package for the statistical programming language R that enables simultaneous analysis of numerous bioassay experiments performed across diverse compounds and biological targets.[20] bioassayR is distinct from existing tools for analyzing high throughput screening data in several important ways: (i) its focus on the simultaneous tracking and comparative analysis of a large number of assays of distinct experimental design and source; (ii) its flexible data structures optimized for performance with large data and interoperability with existing statistical software; (iii) its integration with numerous R language cheminformatics and bioinformatics tools curated by the Bioconductor and CRAN projects, including ChemmineR, ChemmineOB, rcdk, cellHTS, fmcsR, and eiR.[20−26] For example, users can analyze their own HTS data (e.g., processed with cellHTS) alongside public bioactivity data; or process bioactivity fingerprints (HTSFPs) with functionalities provided by ChemmineR. HTSFPs summarize the activity of compounds across many protein targets. Several studies have demonstrated their effectiveness in predicting and categorizing bioactivity in a manner complementing rather than overlapping with structure based predictions.[27−33] In addition,

Received: February 24, 2016
Published: July 1, 2016

DOI: 10.1021/acs.jcim.6b00109
J. Chem. Inf. Model. 2016, 56, 1237−1242

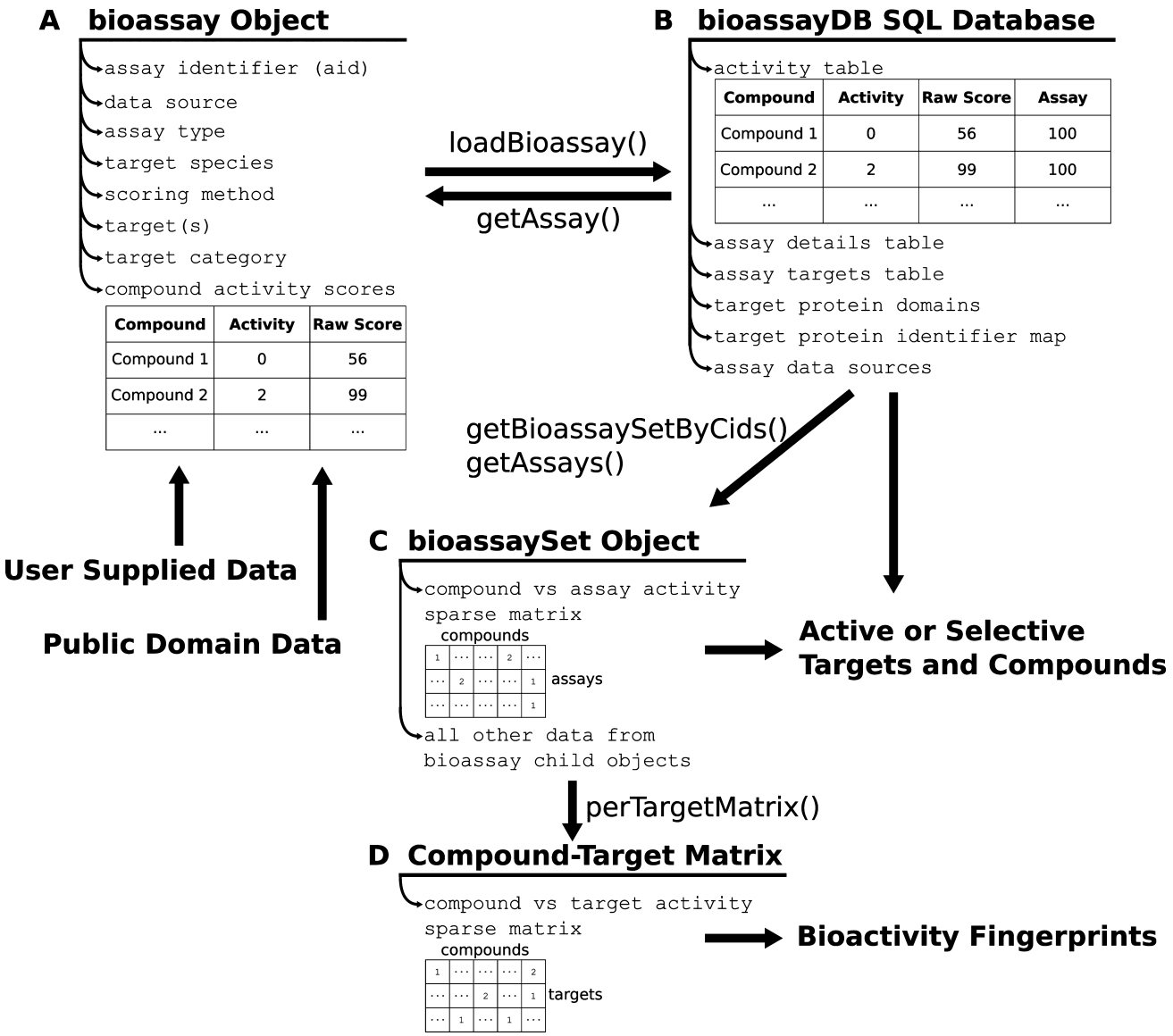


Figure 1. Design overview and workflow. bioassayR stores bioactivity data in four interconnected objects. (A) Data from a single bioassay experiment is imported into a bioassay object. (B) Any number of bioassay objects can be loaded into the *bioassayDB* SQL database that is optimized for time efficient searching. (C) Filter and query methods are available to identify compounds or assays of interest. These query results can be imported into a bioassaySet object that stores activity data as a sparse matrix where columns represent compounds and rows assays (targets). This organization facilitates many typical cross-target analysis routines, e.g., target selectivity analyses. (D) To reduce both redundancy and sparseness in the data, assays involving the same or similar targets can be collapsed into a single row using the *perTargetMatrix* function.

they can be used as trainings data sets for predicting active ligand-target pairs with supervised machine learning algorithms.[34,35] The HTSFP tools implemented in bioassayR will generate fingerprints for any custom set of compounds and targets, optionally merge assays with similar or identical targets, and compare activity profiles by either continuous *z*-scores or binary active/inactive values. The *z*-score based HTSFPs exhibit greater predictive power in hit expansion experiments, whereas binary HTSFPs require less computational overhead, enabling all-against-all bioactivity profile comparison for hundreds of thousands of compounds.[30]

METHODS AND IMPLEMENTATION

Software Design and Workflow Overview. bioassayR's data model is designed around four interconnected data objects (R language S4 classes), each with an internal structure optimized for different bioactivity analysis routines. They are introduced below in more detail, and Figure 1 provides an illustration. In short, the bioassay data is organized in an SQL database called *bioassayDB*; data from single and many assays are imported into *bioassay* and *bioassaySet* objects, respectively; and the compound−target matrix summarizes the compound vs target activities from many assays. The *bioassayDB* serves as a large data repository that can efficiently organize and query millions of assays simultaneously, whereas the other objects facilitate analysis of a subset of these data selected to answer a specific biological question. Table S2 in the Supporting Information lists selected cross-target analysis functions that query the data within these objects. Users can optionally use a prebuilt *bioassayDB* database that contains publicly available bioactivity data against a wide range of protein targets.

DOI: 10.1021/acs.jcim.6b00109
J. Chem. Inf. Model. 2016, 56, 1237−1242

bioassay Object: Importing Data. The *bioassay* object (Figure 1 section A) stores data from a single bioassay experiment, and acts as a gateway for importing new assay data, as well as for editing and investigating data from one assay at a time. This object stores the assay identifier (aid), data source, assay type, target species, scoring method, target identifiers, target categories, and activity scores.

bioassayR provides users with the option of performing analyses either on their own bioactivity data, on a prebuilt database of public domain bioactivity data, or both simultaneously. Four options exist for importing data as a bioassay object: (i) data in the standard PubChem CSV and XML formats can be parsed with a built in function; (ii) data already represented as an R *data.frame* or tabular file with activity values can be directly converted into a bioassay object; (iii) raw screening data from a microtiter plate reader can be analyzed using the cellHTS2 R package, and converted into a bioassay object; (iv) extracting a single assay from an already existing bioassayR database, such as the prebuilt PubChem BioAssay database described below.[26] All four options are demonstrated with examples in the package documentation. Once represented as a bioassay object, these data can be viewed, edited, or loaded into a *bioassayDB* database for analysis alongside other assays.

bioassayDB Object: Multiple Assay SQL Database. The *bioassayDB* object (Figure 1 section B) stores a connection to a SQL database optimized for efficient aggregate search-based analysis across multiple assays. Users can load, edit, or delete individual *bioassay* objects, and then query these data. Many analysis and query functions are provided to investigate the data within a *bioassayDB* object (see Table S2 in the Supporting Information). The database is contained within a single file that can be easily shared among users. Internally, the database stores data from a large number of individual bioassay objects, in addition to target protein domain data, and target identifier mappings. Multiple types of identifier mapping and annotation data can be stored, for example to translate target identifiers into those used by common databases such as UniProt, or to annotate proteins by storing categorization data such as a sequence-similarity clustering bin for each protein.[36]

bioassaySet Object: Storing Multiple Assays in a Matrix. Query results from *bioassayDB* can be stored as a *bioassaySet* (Figure 1 section C). This matrix-like object along with its accessor methods abstracts complicated analysis tasks across large numbers of compounds and bioassays. By representing bioactivity data as a compound vs assay matrix, the full range of matrix operations in R can be leveraged to analyze these data efficiently. For example, rows can be compared to compute the similarity between the activity profiles of two molecules. Sparse matrix compression is utilized to avoid unnecessary usage of system memory by untested compound−target combinations. In a typical workflow, a user will first query the database to find a list of compounds or assays of interest, and then extract these into a *bioassaySet* for further analysis.

To address questions of compound vs target bioactivity, bioassayR can transform a *bioassaySet* into a compound−target matrix by merging assays that share common or similar target proteins, such as close orthologs from different species. Replicates and similar-target assays can be summarized into single values by either specifying a custom summary statistic, or choosing among several provided. The compound vs target matrix can be generated from either discrete "active" or "inactive" activity categories, or from continuous activity scores to serve as either binary or continuous numeric HTSFPs, respectively. The *scaleBioassaySet* function will scale and center continuous scores to create a *z*-score fingerprint. Optionally, omitting inactive values from the discrete activity categories will produce a matrix suitable for analysis with binary matrix algorithms. This data structure can serve as a bipartite graph (or bigraph) connecting compounds and targets, allowing users to analyze these data with the numerous graph and network analysis algorithms available for the R programming language.

Prebuilt PubChem BioAssay Database. To enable efficient analyses across large numbers of compounds and protein targets, we provide downloadable instances of the *bioassayDB* database preloaded with public bioactivity data. This frequently updated database file includes all screens from PubChem BioAssay involving known target proteins. PubChem BioAssay data has been chosen since it includes assays from many sources such as ChEMBL, and therefore represents a substantial portion of all publicly available bioactivity data. At the time of this writing the data contains activity results from roughly 1.2 million structurally distinct compounds tested against protein 6339 targets. As many compound−target combinations have not been tested, these data are sparse with roughly half (572 947) of the compounds having screening results for at least 10 distinct protein targets. Among these "highly screened" compounds, 895 are currently FDA approved drugs. PubChem BioAssay provides bioactivity data both as continuous numeric scores, and active/inactive categories.

To extend the utility of these data, we provide and include within the prebuilt database additional annotation details for each protein target. The database includes both NCBI Protein GI numbers and UniProt identifiers for all protein targets, Pfam domains identified with the HMMER software, and amino acid sequence similarity-based clustering performed with kClust.[36−39] The UniProt identifiers allow users to obtain further annotation details including Gene Ontology (GO) terms programmatically by connecting to external annotation databases.[40] The Pfam domain mappings provide groupings for local similarities and across wider evolutionary distances, whereas the sequence similarity cluster are more suitable for identifying groups of sequences sharing a defined degree of sequence similarity.

The included annotation data expand the usefulness of bioassayR for several applications. For instance, the annotations can be used for merging similar assays into a compound−target matrix as described in the above "bioassaySet Object" section. When searching for compounds active against a desired protein, users can expand the search to include compounds found active against protein targets that share sequence similarity, domains, or GO terms with the query. This method can identify compounds that are likely active against a target of interest, even if little or no screening data exists for that specific target. In drug discovery experiments where a specific protein target has not yet been identified, these data can help identify protein targets worth investigating based on presence of a specific protein domain, molecular function, or orthologue that has been previously found to be involved in the desired therapeutic effect.

Identifying Compounds with Selective or Promiscuous Bioactivity. Bioactive small molecules can be classified according to the quantity of distinct molecular targets they are active against. Target selective compounds bind to a small number of target proteins, whereas "promiscuous binders",

DOI: 10.1021/acs.jcim.6b00109
J. Chem. Inf. Model. 2016, 56, 1237−1242

indiscriminately bind to a large number of targets. Patterns of target selectivity in widely used drugs can also be used as a template for identifying drug candidates with similar selectivity profiles.

Several bioassayR functions facilitate identification of target selective compounds and the reverse, compound selective targets, across a large set of bioassay experimental results. The *targetSelectivity* function will return the target selectivity for a query compound. To find compounds active against a target or a set of targets in a pathway of interest, the function *activeAgainst* will return all active compounds, whereas *selectiveAgainst* will return only compounds most selective against the specified target, along with a corresponding selectivity score for each. To consider only compounds that have been tested in numerous assays, the *screenedAtLeast* function will identify compounds that have participated in a specified minimum quantity of screens. To find all targets of a query compound, the functions *activeTargets* and *inactiveTargets* will return the list of active and inactive targets, respectively. The *crossReactivityProbability* function uses a beta-binomial statistical model to estimate the probability that a given compound is a promiscuous binder.[13]

Clustering Small Molecules by Bioactivity Profile. With bioassayR, large-scale screening data can be used to cluster small molecules based on the similarity of their bioactivity profiles across many target proteins. To cluster small molecules by bioactivity, it is necessary to choose an appropriate similarity measure, such as correlation coefficients that are appropriate for continuous activity data, and the Jaccard or Tanimoto coefficient for categorical or binary data.[41] Next, the chosen similarity measure is used to compute a distance matrix (d) for all possible pairwise comparisons of bioactivity profiles, by subtracting the similarity values (s) from one: $d = 1 - s$. The distance matrix can then be used as direct input to a variety of clustering algorithms, including hierarchical clustering, k-means or multiple dimensional scaling (MDS).

The bioassayR clustering workflow starts by generating a compound−target bioactivity matrix, as described above, with either continuous or discrete category activity scores. For continuous scores, several similarity functions available in R, such as the base function *cor* can be used to create a distance matrix based on Pearson correlation coefficients. The associated ChemmineR package will create a distance matrix for binary bioactivity fingerprints generated by bioassayR. Comparisons among binary ChemmineR fingerprints have less CPU and memory overhead than continuous z-score based comparisons, and therefore are suitable for all-against-all comparisons of larger compound sets.

By default, the bioassayR HTSFPs features resolve missing (untested) activity values by assuming inactivity, where a "0" is used for binary fingerprints, and a z-score of "0" is used for continuous fingerprints. When computing the similarity between two compound bioactivity profiles, this can lead to false negatives (lower than the true similarity value) if the compounds share few common screened targets.[30] A more accurate estimate of similarity can be obtained by using machine learning methods that impute the missing values; however, this introduces false positives that are often less desirable than false negatives in drug discovery efforts.[30] The bioassayR function *screenedAtLeast* can limit false negatives without introducing false positives by including only highly screened compounds in the analysis. Alternatively, the compound vs target matrix can be subset with a biclustering algorithm to limit similarity comparison to a densely screened subset of a larger sparse compound vs target matrix. Lastly, bioassayR also provides a similarity function (*trinarySimilarity*) that avoids assuming inactivity for missing compound−target activity values by operating on a trinary bioactivity matrix that uses a "0" for untested or missing values, a "1" for inactive values, and a "2" for active values. This function computes similarity based only on the mutually screened targets between two compounds, and returns an "NA" if insufficient shared assays exist to make a meaningful comparison. The strategy of performing the comparison only on mutually screened targets, with a minimum threshold for informative data was inspired by the continuous score "Assay Performance Profile Similarity" metric published by Dančík et al.[13]

RESULTS AND DISCUSSION

In the Supporting Information, we highlight three example use cases demonstrating the utility of bioassayR. First, we investigate the diversity of public screening data provided by PubChem BioAssay, and show that these data contain compounds active against a large number of novel protein targets that are not currently accessible with FDA approved drugs. Second, we use bioassayR to cluster FDA approved drugs by bioactivity profiles as well as molecular structure to demonstrate that many drugs exhibit distinct bioactivity patterns that cannot be inferred from structure alone. Third, we demonstrate how bioassayR can be used to enrich a screening library with active compounds and how to guide the time-consuming target site identification processes in high-content screening. The vignette (user manual) of the package contains additional examples including loading custom screening data, identifying target selective compounds, and performing custom database queries.

It is important to point out that HTS data are noisy and error prone due to several causes including experimental noise, and incorrect annotation. Although public bioactivity databases have implemented strategies to identify and reduce errors, we caution bioassayR users to expect some level of error and mis-annotation depending on the source and type of data used.[42] The impact of these errors on analysis results can be minimized by incorporating replicates and confirmatory screening results from different sources using the bioassayR functions described above. If appropriate, error can also be reduced by limiting analysis to the subset of public bioactivity data that has been manually curated and carefully annotated with a machine readable, nonambiguous structured vocabulary from sources such as the BioAssay Research Database (BARD).[11,43]

The bioassayR package is a flexible computational environment for simultaneous analysis of large numbers of high-throughput small molecule bioactivity screens. By organizing large bioactivity data for rapid access and manipulation within the R programming language, bioassayR leverages the substantial breadth of these data as a reference to identify regions of the genome and proteome accessible to small molecule probes, elucidate mechanisms of action for bioactive molecules, and identify off-target effects that currently lead to a high attrition rate in drug discovery efforts.[44] bioassayR provides features to inform the design and analysis of bioactivity and drug discovery experiments; for example to build compound libraries enriched for a desired bioactivity, reducing the search space for effective drugs, druggable protein targets, and chemical genetic probes. bioassayR has functions to identify compounds that have demonstrated activity against

DOI: 10.1021/acs.jcim.6b00109
J. Chem. Inf. Model. 2016, 56, 1237−1242

targets and pathways of interest, or other targets with sequence or annotation similarity to targets of interest. To build drug discovery libraries with reduced chances of off-target effects, bioassayR will rank compounds for selectivity against a desired target and exclude compounds that show activity against a large number of other targets. To identify compounds or combinations of compounds likely to exhibit a desired polypharmacology (activity against multiple targets), bioassayR will identify all active compounds among a set of query targets. To assess the potential druggability of protein targets, bioassayR will report the quantity and target selectivity of known active drugs and other compounds. To identify compounds with activity similar to existing drugs or other compounds with a known utility, the HTSFP features enable clustering by cross-target activity profiles. Custom screening data can also be analyzed side-by-side with public data to study the selectivity profiles among newly identified actives across numerous targets, or to assess the level of agreement with any public data that the custom assay replicates.

In addition to providing numerous analysis functions, bioassayR also serves as a bridge to facilitate analysis of large screening data with other machine learning, statistical inference, network analysis, and bioinformatics tools. Many of these tools support the output formats produced by bioassayR with little or no changes. In conclusion, bioassayR lowers the barrier to address questions related to the target selectivity of small molecules with large-scale bioactivity data.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00109.

Table of small molecule databases, cross-target analysis functions, use case examples, and a performance evaluation (PDF).

■ AUTHOR INFORMATION

Corresponding Author

*T. Girke. E-mail: thomas.girke@ucr.edu.

Funding

This project was supported by grants from the National Science Foundation [ABI-0957099] and the National Institute of Health [U24AG051129].

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

HTSFPs, high throughput screening fingerprints; HTS, high throughput screening

■ REFERENCES

(1) 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.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.

(2) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(3) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(4) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21*, 4133−4139.

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

(6) Yan, S. F.; King, F. J.; He, Y.; Caldwell, J. S.; Zhou, Y. Learning from the Data: Mining of Large High-Throughput Screening Databases. *J. Chem. Inf. Model.* **2006**, *46*, 2381−2395.

(7) Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational Analysis of Multi-target Structure-Activity Relationships to Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem* **2010**, *5*, 847−858.

(8) Han, L.; Wang, Y.; Bryant, S. H. A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics* **2009**, *25*, 2251−2255.

(9) Cheng, T.; Wang, Y.; Bryant, S. H. Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules. *Bioinformatics* **2010**, *26*, 2881−2888.

(10) Senger, C.; Gruning, B. A.; Erxleben, A.; Doring, K.; Patel, H.; Flemming, S.; Merfort, I.; Gunther, S. Mining and evaluation of molecular relationships in literature. *Bioinformatics* **2012**, *28*, 709−714.

(11) Schurer, S. C.; Vempati, U.; Smith, R.; Southern, M.; Lemmon, V. BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. *J. Biomol. Screening* **2011**, *16*, 415−426.

(12) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.* **2003**, *46*, 4265−4272.

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

(14) Schmidt, U.; Struck, S.; Gruening, B.; Hossbach, J.; Jaeger, I. S.; Parol, R.; Lindequist, U.; Teuscher, E.; Preissner, R. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.* **2009**, *37*, D295−D299.

(15) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **2010**, *6*, 343.

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

(17) Dandapani, S.; Marcaurelle, L. A. Grand Challenge Commentary: Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.* **2010**, *6*, 861−863.

(18) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138−1139.

(19) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924−1936.

(20) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.

(21) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y. H.; Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.

(22) Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T. ChemmineR: a compound mining framework for R. *Bioinformatics* **2008**, *24*, 1733−1734.

(23) Guha, R. Chemical Informatics Functionality in R. *J. Stat. Softw.* **2007**, *18*, 1−16.

DOI: 10.1021/acs.jcim.6b00109
J. Chem. Inf. Model. 2016, 56, 1237−1242

(24) Cao, Y.; Jiang, T.; Girke, T. Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. *Bioinformatics* **2010**, *26*, 953−959.
(25) Wang, Y.; Backman, T. W. H.; Horan, K.; Girke, T. fmcsR: mismatch tolerant maximum common substructure searching in R. *Bioinformatics* **2013**, *29*, 2792−2794.
(26) Boutros, M.; Brás, L. P.; Huber, W. Analysis of cell-based RNAi screens. *Genome Biol.* **2006**, *7*, R66.
(27) Helal, K. Y.; Maciejewski, M.; Gregori-Puigjané, E.; Glick, M.; Wassermann, A. M. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J. Chem. Inf. Model.* **2016**, *56*, 390−398.
(28) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399−1409.
(29) Wassermann, A. M.; Lounkine, E.; Urban, L.; Whitebread, S.; Chen, S.; Hughes, K.; Guo, H.; Kutlina, E.; Fekete, A.; Klumpp, M.; Glick, M. A Screening Pattern Recognition Method Finds New and Divergent Targets for Drugs and Natural Products. *ACS Chem. Biol.* **2014**, *9*, 1622−1631.
(30) Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1880−1891.
(31) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol. (Oxford, U. K.)* **1995**, *2*, 107−118.
(32) Kauvar, L. M.; Villar, H. O.; Sportsman, J. R.; Higgins, D. L.; Schmidt, D. E., Jr Protein affinity map of chemical space. *J. Chromatogr., Biomed. Appl.* **1998**, *715*, 93−102.
(33) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 261−266.
(34) Lusci, A.; Browning, M.; Fooshee, D.; Swamidass, J.; Baldi, P. Accurate and efficient target prediction using a potency-sensitive influence-relevance voter. *J. Cheminf.* **2015**, *7*, 63.
(35) Soufan, O.; Ba-alawi, W.; Afeef, M.; Essack, M.; Rodionov, V.; Kalnis, P.; Bajic, V. B. Mining Chemical Activity Status from High-Throughput Screening Assays. *PLoS One* **2015**, *10*, e0144426.
(36) The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75; DOI: 10.1093/nar/gkr981.
(37) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.; Finn, R. D. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290−D301.
(38) Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **2009**, *23*, 205−211.
(39) Hauser, M.; Mayer, C. E.; Söding, J. kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinf.* **2013**, *14*, 248.
(40) The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **2015**, *43*, D1049−D1056; DOI: 10.1093/nar/gku1179.
(41) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.
(42) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 885−896.
(43) Howe, E. A.; de Souza, A.; Lahr, D. L.; Chatwin, S.; Montgomery, P.; Alexander, B. R.; Nguyen, D.-T.; Cruz, Y.; Stonich, D. A.; Walzer, G.; Rose, J. T.; Picard, S. C.; Liu, Z.; Rose, J. N.; Xiang, X.; Asiedu, J.; Durkin, D.; Levine, J.; Yang, J. J.; Schurer, S. C.; Braisted, J. C.; Southall, N.; Southern, M. R.; Chung, T. D. Y.; Brudz, S.; Tanega, C.; Schreiber, S. L.; Bittker, J. A.; Guha, R.; Clemons, P. A. BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.* **2015**, *43*, D1163−D1170.
(44) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discovery* **2012**, *11*, 909−922.

DOI: 10.1021/acs.jcim.6b00109
J. Chem. Inf. Model. 2016, 56, 1237−1242