PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Predicting kinase inhibitors using bioactivity matrix derived informer sets

Huikun Zhang[1☯], Spencer S. Ericksen[2☯], Ching-pei Lee[3☯], Gene E. Ananiev[2], Nathan Wlodarchak[4], Peng Yu[1], Julie C. Mitchell[5], Anthony Gitter[6,7], Stephen J. Wright[8], F. Michael Hoffmann[2,9], Scott A. Wildman[2], Michael A. Newton[1,6]*

1 Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **2** Small Molecule Screening Facility, Drug Development Core, UW-Carbone Cancer Center, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **3** Department of Mathematics and Institute for Mathematical Science, National University of Singapore, Singapore, **4** Department of Medicine, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **5** Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennesse, United States of America, **6** Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **7** Morgridge Institute for Research, Madison, Wisconsin, United States of America, **8** Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **9** Department of Oncology, McArdle Laboratory for Cancer Research, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

☯ These authors contributed equally to this work.
* newton@stat.wisc.edu

Citation: Zhang H, Ericksen SS, Lee C-p, Ananiev GE, Wlodarchak N, Yu P, et al. (2019) Predicting kinase inhibitors using bioactivity matrix derived informer sets. PLoS Comput Biol 15(8): e1006813. https://doi.org/10.1371/journal.pcbi.1006813

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: January 24, 2019

Accepted: July 13, 2019

Published: August 5, 2019

Data Availability Statement: All data used in our manuscript are available at (https://github.com/SpencerEricksen/informers/).

Funding: MAN, FMH, SAW, SSE were supported in part by the core grant of the UW Comprehensive Cancer Center; NCI: P30 CA014520. MAN, SSE, HZ, and AG were supported in part by NIH U54AI117924, a grant supporting the UW Center for Predictive Computational Phenotyping. SJW and MAN were supported in part by NSF CCF 1740707. NJW was supported in part by the American Heart Association 19POST34380404.

Abstract

Prediction of compounds that are active against a desired biological target is a common step in drug discovery efforts. Virtual screening methods seek some active-enriched fraction of a library for experimental testing. Where data are too scarce to train supervised learning models for compound prioritization, initial screening must provide the necessary data. Commonly, such an initial library is selected on the basis of chemical diversity by some pseudo-random process (for example, the first few plates of a larger library) or by selecting an entire smaller library. These approaches may not produce a sufficient number or diversity of actives. An alternative approach is to select an informer set of screening compounds on the basis of chemogenomic information from previous testing of compounds against a large number of targets. We compare different ways of using chemogenomic data to choose a small informer set of compounds based on previously measured bioactivity data. We develop this Informer-Based-Ranking (IBR) approach using the Published Kinase Inhibitor Sets (PKIS) as the chemogenomic data to select the informer sets. We test the informer compounds on a target that is not part of the chemogenomic data, then predict the activity of the remaining compounds based on the experimental informer data and the chemogenomic data. Through new chemical screening experiments, we demonstrate the utility of IBR strategies in a prospective test on three kinase targets not included in the PKIS.

The team was also supported by an internal UW2020 grant from the University of Wisconsin Madison, Office of the Vice Chancellor for Research and Graduate Education and the Wisconsin Alumni Research Foundation (https://research.wisc.edu/funding/uw2020/round-3-projects/). Computations were supported in part by NSF award 1148698 to the UW Center for High-Throughput Computing and Open Science Grid and also by the UW Biostatistics Computing Group. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

In the early stages of drug discovery efforts, computational models are used to predict activity and prioritize compounds for experimental testing. New targets commonly lack the data necessary to build effective models, and the screening needed to generate that experimental data can be costly. We seek to improve the efficiency of the initial screening phase, and of the process of prioritizing compounds for subsequent screening.

We choose a small *informer set* of compounds based on publicly available prior screening data on distinct targets. We then collect experimental data on these informer compounds and use that data to predict the activity of other compounds in the set for the target of interest. Computational and statistical tools are needed to identify informer compounds and to prioritize other compounds for subsequent phases of screening. We find that selection of informer compounds on the basis of bioactivity data from previous screening efforts is superior to the traditional approach of selection of a chemically diverse subset of compounds. We demonstrate the success of this approach in retrospective tests on the Published Kinase Inhibitor Sets (PKIS) chemogenomic data and in prospective experimental screens against three additional non-human kinase targets.

Introduction

Early-stage drug discovery involves a search for pharmacologically active compounds (hits) that produce a desired response in an assay of protein function or disease-related phenotype. The active compounds serve as starting points for further structural optimization, with the ultimate goal of developing therapeutic agents. Virtual screening (VS) can be an effective strategy for prioritizing compounds that can lower high-throughput screening costs by reducing the experimental search to smaller, active-enriched compound subsets. This process can be cheaper and more effective than exhaustive, unguided testing of entire compound libraries [1]. VS may also allow us to evaluate much larger physical or virtual compound libraries. As on-demand synthetic capabilities expand, a VS-guided approach might obviate costs associated with purchasing and on-site storage/maintenance of large general libraries in favor of growing smaller, project-focused compound sets [2].

The choice of which VS methodology to deploy depends on the types of information available at the start of this effort [3]. Structure-based VS methods (such as docking) require specific, structurally-characterized biomolecular targets, but these target structures might only be approximated by homology models [4], or might not be available at all. Phenotypic endpoints like cell death or tumor shrinkage are not amenable to structure-based approaches because specific target structures and sites of action may not be known. Furthermore, structure-based VS performance varies substantially across targets, where failures are difficult to predict [4, 5]. Ligand-based VS approaches can provide more consistent levels of enrichment and are independent from any target structure, but they depend strongly on the quality and abundance of training data in the form of measured compound activities on the target of interest [6]. Such approaches, especially those using topological features for compound representations (such as graph-based fingerprints), may also suffer from high prediction uncertainty when presented with compounds whose chemotypes/scaffolds are outside the scope of the training set [6, 7]. The key issue, however, is that training data are usually scarce in early stages of the screening process, making it difficult to generate a predictive model.

For some well-studied target classes (for example, kinases or GPCRs), rich chemogenomic data are available in the form of compound activity profiles across many members of a target class. These data can be structured as a targets-by-compounds matrix of functional interactions, which we term the *bioactivity matrix*. Though sometimes sparse, incomplete, or limited in compound and target coverage, such matrices hold valuable information that can be leveraged to make predictions on new targets or compounds.

Predictions of compound activities are routinely made using machine learning algorithms to relate a selection of chemical features to the previously measured bioactivities of a training set of compounds. In many cases, these features are chemical fingerprints that describe the presence and proximity of chemical substructures in each compound [6, 7]. Alternative compound fingerprints have been developed on the basis of prior chemogenomic data [8–13]. In these cases, the bioactivity profile of a compound across a series of assays is used as a fingerprint, referred to as a "High Throughput Screening FingerPrint" (HTS-FP), based either on continuous bioactivity values or on a binary quantity representing activity/inactivity. HTS-FPs enable a useful expression of compound relationships through distances derived among standardized bioactivity profiles in much the same manner as chemical fingerprints. HTS-FPs have limited extensibility in that the wide array of assays/target responses that confers a rich pharmacological representation cannot be readily generated for new molecules. However, looking beyond compound representations, arrays of standardized bioactivity data, even when incomplete, can help to establish target relationships.

Given a new target with little or no prior structure–activity relationship information, building an effective ligand-based VS model requires training data acquired through preliminary screening. For the virtual screening model to be cost effective, the library subset providing training instances should be as small as possible. However, preliminary unguided screens constrained to only 100s to 1000s of compounds are likely to produce insufficient training data with few active training instances of limited potency and structural diversity.

Motivated by the need for batch selection strategies to enable effective iterative screening efforts, there has been significant recent effort in developing compound prioritization models from minimal data [14–18]. These methods prioritize additional compounds for testing based on an initial increment of screening data, but the selection of the initial subset of compounds to be screened is often random, pseudo-random, or based on chemical diversity. A recent effort by Paricharak et al. [18] uses an active learning process to select an informer set from the most active and least active compounds across a series of PubChem assays. Their work removes specific assay labels from the chemogenomic data to create a balanced data set, and selects compounds on the basis of uncertainty from previous predictive models. However, their optimal informer set is too large to be useful as an initial screening set in most HTS settings.

Our emphasis in this paper is on the selection of the *informer set*—the initial set of compounds to be assayed. The experimental data for these compounds may then be used to train initial models or to select additional compounds as the initial (0^{th} iteration) set of compounds to be assayed in a multi-phase scheme. We refer to approaches based on informer sets as *Informer-Based Ranking* (IBR) methods. This is different than the focus of the studies cited above that focus on model-guided or heuristic selection of compounds for multiple phases of screening. Our approaches are analogous to earlier chemometric experimental design approaches like chemical cluster sampling [19], but leverage chemogenomic data instead. The proposed IBR methods each involve two steps; see Fig 1. In the first step, they select an informer set of compounds to evaluate experimentally for bioactivity on the new target. Importantly, this selection is guided by the bioactivity matrix. The second step involves prioritization of the compounds outside the informer set, according to their bioactivity against the

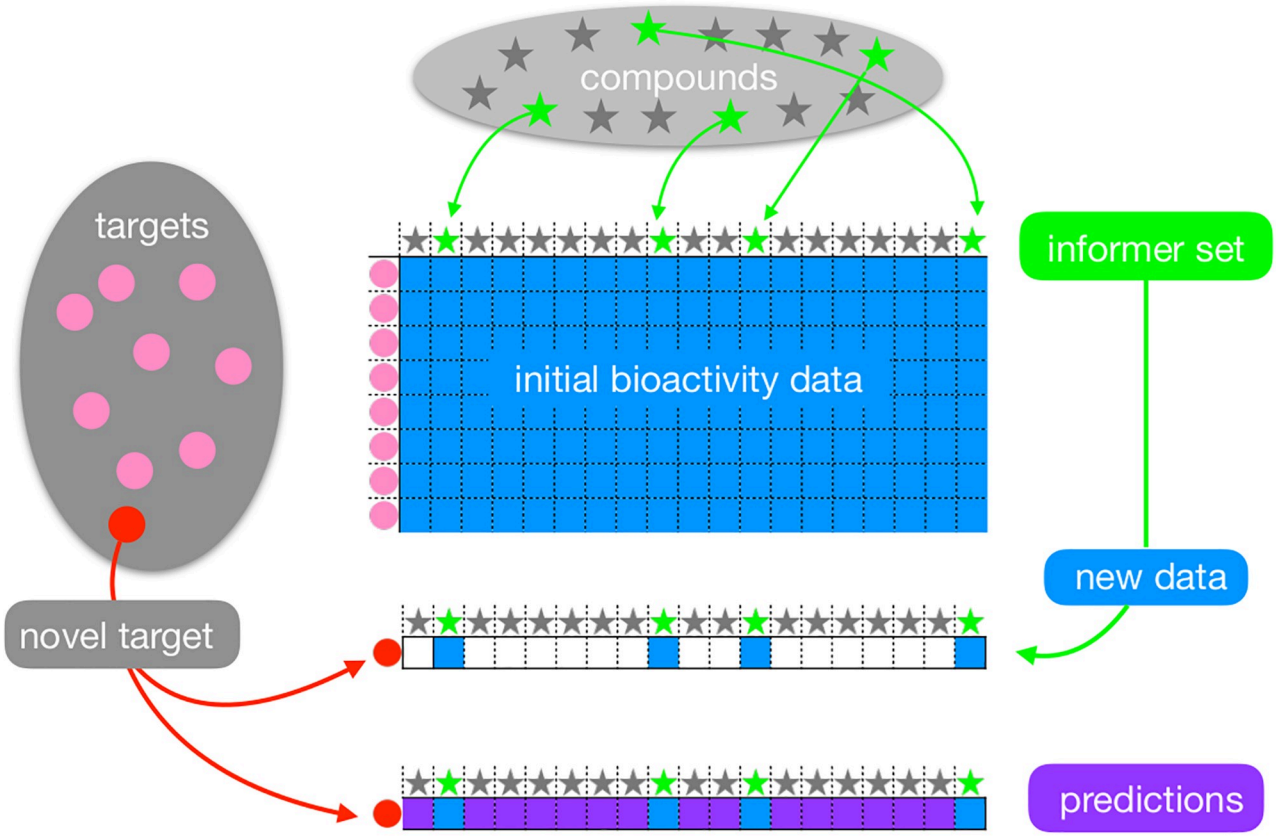

Fig 1. IBR (Informer-Based Ranking) for compound prioritization on a novel target. From a complete bioactivity data matrix (blue grid), a subset of informer compounds (green stars) are identified from the broader set of compounds (stars) that have been tested against a large set of targets (pink circles). A previously uncharacterized target (red circle) is assayed with just the informer compounds, and the new bioactivity data are used to reveal the new target's relationship to other targets. The combined data enable activity predictions (purple) on the remaining, non-informer compounds.

https://doi.org/10.1371/journal.pcbi.1006813.g001

new target. This prioritization may make use of both the bioactivity data on other targets as well as the new data obtained on the informer compounds on the new target. We describe algorithms both for the selection of the informer compounds and for the prioritization of other compounds after screening data are obtained for the informer compounds.

We propose three novel IBR strategies: Regression Selection (RS), Coding Selection (CS), and Adaptive Selection (AS). Each strategy consists of (i) an informer set selection method that chooses a small number of compounds to be tested, based only on characteristics of the bioactivity matrix, and (ii) a compound ranking method that leverages returned informer data to predict which of the untested compounds is active against a new target. Underlying all three strategies is the premise that targets may be naturally organized according to patterns in their bioactivity profiles across compounds. This organization leads to a clustering of targets as well as to the identification of informer compounds that are predictive of the cluster identity of a novel target. The strategies leverage advances in optimization and statistical analysis, and they differ in how patterns are recognized and computations are deployed.

We apply the proposed IBR strategies in the context of two public human kinase chemogenomics matrices: PKIS1 [20] and PKIS2 [21]. We demonstrate the strategies prospectively, by prioritizing PKIS1 and PKIS2 compounds for activity against three distinct protein kinase targets of potential therapeutic importance: *Mycobacterium tuberculosis* PknB [22], Epstein-Barr virus BGLF4, and *Toxoplasma gondii* ROP18 [23]. We also apply the strategies retrospectively,

in a cross-validation study of each chemogenomics matrix, leaving out one target at a time and prioritizing compound activity against the left-out target.

The performance of each new IBR strategy was assessed prospectively by inspection of the successful activity predictions, and retrospectively using common VS metrics, including: Area Under the Receiver-Operator Characteristic Curve (ROCAUC) and enrichment factor (EF). We also assessed each strategy's ability to retrieve structural diversity among active compounds by computing the fraction of active scaffolds identified in the top of the ranking. For benchmarking purposes, we compared the proposed IBR methods to a set of baseline models that make use of the compound structures, and include the commonly used diverse selection as well as a selection of the most frequently active compounds in the bioactivity matrix.

Results

IBR strategies apply in the low-data regime

We describe IBR strategies that require experimental testing of some new target of interest on a small fraction of the compound library—the informer subset—with a view to effectively prioritizing the remaining compounds for subsequent testing for activity with the target. The complete IBR strategy thus has two parts: a scheme to identify the informer subset and a scheme to prioritize the remaining compounds after assay data have been obtained for the informer compounds. Initially, we may have no assay data on the new target, though we typically have some such chemogenomic data on related targets that populate a related sector of chemical space, in some sense. Ideally, a successful IBR strategy might be applied in target-agnostic drug development settings (for example, phenotypic targets or incompletely featurized targets), so we intentionally exclude from each IBR strategy target-specific features, such as protein sequences or structural information.

We described three novel IBR strategies that use statistical patterns in the bioactivity matrix that is available prior to informer-set assay testing. Regression Selection (RS), Coding Selection (CS), and Adaptive Selection (AS) all treat the target space as being partitioned into clusters of targets so that, within each cluster, there is some relevant similarity of the bioactivity profiles of the targets across the space of tested compounds. These three strategies also posit that a small number of compounds (the informer subset) have bioactivity profiles that are predictive of the cluster label appropriate to any target, including the novel target of interest. RS, CS, and AS differ in how they evaluate clusterings and potential informer subsets. For example, RS and AS involve `kmeans` clustering of targets followed by regularized multinomial regression to learn the relationship between compounds and cluster labels, but they differ in how the regression is regularized and how the informer compounds are identified. In contrast, CS forms a single objective function that simultaneously scores clustering strategies and potential informer compounds.

Computationally simpler baseline IBR strategies are useful to consider, as they may approximate practical experimental design scenarios. Baseline Chemometric strategies (BC_s, BC_l, and BC_w) use chemical features for both informer selection and non-informer ranking. Three different chemometric ranking strategies are used for the non-informer ranking, as denoted by subscripts s, l, and w (described in detail in the Methods). Here, clustering is applied on the compound space using the known chemical structure (fingerprints) of the compounds (not used in RS, CS, or AS) in order to identify informer compounds. Then, prioritization of the non-informers makes use of various ways of ranking the chemical distance between bioactive informers and non-informers. Alternatively, a Baseline *Frequent-hitters* strategy simply takes as informer compounds those that show the highest rate of activity within the initial target set (BF_s, BF_l, BF_w). Prioritization of non-informers uses chemical distance, as in the chemometric

methods. To simplify, we only report baseline results for each of our top chemometric and frequent-hitters baseline strategies (BC_w and BF_w). Outcomes for the full set of baselines are available in the supplemental information.

Performance of the IBR strategies was evaluated using two virtual screening metrics that reflect successful prioritization of active compounds: ROCAUC and Normalized Enrichment Factor in top 10% of ranking (NEF10)). An additional metric Fraction of Active Scaffolds Retrieved (FASR10) assesses the diversity of the active chemical structures that were prioritized in the top 10% of the ranking. Also, standard classification metrics F1 score and MCC were applied.

Prospective tests of IBR strategies on novel kinase targets

We applied the IBR strategies on three novel kinase targets outside of the PKIS1 and PKIS2 target sets. These microbial targets are phylogenetically distant from most of the human protein kinases in the PKIS data sets, with relatively low kinase domain sequence identities to the nearest neighbors in the PKIS1/2 sets in comparison to kinase domain sequences (S1 Fig). For *Mycobacterium tuberculosis* kinase PknB (UniProt ID: P9WI81), the nearest neighbors were the human serine/threonine kinases MARK2 (16.1% kinase domain sequence identity) in PKIS1 and BRSK1 (16.1%) in PKIS2 (UniProt IDs: Q7KZI7 and Q8TDC3, respectively). For Epstein-Barr virus kinase BGLF4 (UniProt ID: I1YP37), the most similar kinase domain sequences were from human protein tyrosine kinase (PTK2 or FAK2) (13.8%) in PKIS1 and human serine/threonine-protein kinase (LRRK2) (14.2%) in PKIS2 (UniProt IDs: Q14289.2 and Q5S007). For *Toxoplasma gondii* ROP18 (UniProt ID: Q2PAY2), the most similar kinase domains are NEK7 (UniProt ID: Q8TDX7.1) (20.2%) in PKIS1 and aurora kinase C (AURKC, UniProt ID: Q9UQ89.1) (20.7%) in PKIS2.

To prioritize which PKIS compounds might be active on PknB, BGLF4, or ROP18, each IBR strategy selected 16 informer compounds from PKIS1 and 16 informer compounds from PKIS2. PknB and BGLF4 were obtained and screened in-house while ROP18 data were collected from an external collaborator [23]. The screening data were held separately from the IBR and baseline method operators prior to informer selection. After selecting PKIS1 and PKIS2 informer compounds, screening data only for those compounds were provided to each IBR and baseline method. Informer set selections by each IBR for PKIS1 are shown with their associated experimental bioactivity measurements in S1 Table. The assay results for the informer compounds selected by each of the IBR strategies were used to rank the remaining non-informers in PKIS1 or PKIS2. To evaluate the performance of the different methods, all of the available PKIS1 and PKIS2 compounds were assayed. Experimental active/inactive labels were assigned using $\mu + 2\sigma$ percent inhibition (activity) thresholds in PKIS1: PknB = 13.4%, BGLF4 = 20.2%, and ROP18 = 43.8% and PKIS2: PknB = 8.7%, BGLF4 = 12.5%, and ROP18 = 33.4% based on screening results from the PKIS compound sets.

The RS and CS approaches were the only methods that recovered multiple hits and active scaffolds in their top 10% of ranked compounds for all three kinase targets and both PKIS datasets (Table 1 and S2 Table). RS managed to recover actives for PknB even though it did not include any active compounds in its PKIS1 or PKIS2 informer sets. The RS method was also the best overall for BGLF4 on PKIS2 and tied as the best method for PknB on PKIS1. CS was the best approach for PknB on PKIS2.

AS and the three BF baseline methods (BF_w shown in Table 1) struggled for PknB and BGLF4 with the PKIS2 compounds, each identifying only a single hit. However, AS was the best approach for BGLF4 on PKIS1 compounds and performed better on ROP18 with PKIS2 compounds. The three purely chemometric baseline approaches (BC) (BC_w shown in Table 1)

Table 1. Retrieval counts by the various methods on new kinase targets (a) PknB, (b) BGLF4, and (c) ROP18 using PKIS1 or PKIS2 matrices. The total number of experimentally determined active compounds and distinct active scaffolds is indicated in the *total* column. The values below each of the IBR methods indicate the number of active informers identified, the number of experimentally determined active compounds that were ranked in the top 10% of predicted active compounds by each method, and the number of unique active scaffolds identified in those top 10%. For a given target, these 10% are the active informers and the top ranking non-informers comprising 10% of the set of all compounds after removing inactive informers.

(a) PknB

		baselines		non-baselines			
matrix	hits	BC_w	BF_w	RS	CS	AS	total
PKIS1	active compounds	1	7	7	2	3	8
	active scaffolds	1	7	7	2	3	8
PKIS2	active compounds	0	1	2	3	1	7
	active scaffolds	0	1	2	3	1	7

(b) BGLF4

		baselines		non-baselines			
matrix	hits	BC_w	BF_w	RS	CS	AS	total
PKIS1	active compounds	3	9	3	7	10	11
	active scaffolds	2	6	3	5	7	8
PKIS2	active compounds	1	1	8	3	1	10
	active scaffolds	1	1	7	3	1	8

(c) ROP18

		baselines		non-baselines			
matrix	hits	BC_w	BF_w	RS	CS	AS	total
PKIS1	active compounds	4	7	4	4	2	16
	active scaffolds	3	4	2	3	2	11
PKIS2	active compounds	7	5	3	3	5	19
	active scaffolds	4	3	2	2	3	12

https://doi.org/10.1371/journal.pcbi.1006813.t001

were the worst overall, in many cases failing to recover any hits. Nevertheless, BC_s and BC_l were the top methods on ROP18 with PKIS2 compounds (S2 Table). The best methods were the same when evaluated with NEF10 or FASR10, but varied slightly for ROCAUC (S3 Table).

Retrospective tests of IBR strategies by cross validation on PKIS1 data matrix

The PknB, BGLF4, and ROP18 results demonstrate that the IBR methods perform reasonably well even in a challenging setting where the new targets have low kinase domain similarity with the targets used to construct the informer set. For a more comprehensive quantitative assessment of the IBR methods, we conducted retrospective leave-one-target-out (LOTO) analysis for each of the $m = 224$ targets in PKIS1. This involved $m = 224$ separate applications of all the IBR strategies applied to reduced chemogenomics matrices ($m - 1$ rows), again using an informer size of 16 compounds. Each time, the bioactivity profile of the left-out target was predicted in the sense that compounds were prioritized for activity against this one left-out target.

Results from PKIS1 LOTO cross validation are summarized in Table 2. With respect to the ROCAUC metric (Fig 2), the purely bioactivity-based RS model provides the best rankings with a median ROCAUC value of 0.92 ± 0.11 (± one standard deviation). RS and AS methods both had better performance than the top chemocentric and frequent-hitter baseline approaches, BC_w (0.67 ± 0.22) and BF_s (0.83 ± 0.14). The improvements in ROCAUC of RS and AS over BC_w ($p = 5.5E-31, 2.0E-23$) and BF_s ($p = 1.3E-21, 4.9E-6$) were statistically significant. All p-values were obtained from a 2-sided, pairwise Wilcoxon sign-rank test with Šidák

Table 2. (a) ROCAUC, (b) NEF10, and (c) FASR10 in Leave-One-Target-Out Cross Validation on PKIS1. IBR methods were evaluated on 224 PKIS1 targets using standard VS metrics that reflect active retrieval: ROCAUC and NEF10. FASR10 was also evaluated to reflect the chemical diversity of the actives retrieved. All baseline outcomes are shown in S4 Table along with *p*-values from pairwise comparisons in S5 Table. *The only non-baseline IBR that fails to demonstrate statistical improvement ($p < 0.0085$) over all baselines is CS when using the ROCAUC metric. Note: a Šidák multiple comparison correction was applied using 6 baselines against each non-baseline IBR, lowering the α threshold from 0.05 to 0.0085.

(a) ROCAUC

	baselines		non-baselines		
	BC_w	BF_w	RS	*CS	AS
mean	0.63	0.79	0.90	0.81	0.84
median	0.67	0.81	0.93	0.83	0.88
stdev	0.21	0.13	0.11	0.14	0.14

(b)NEF10

	baselines		non-baselines		
	BC_w	BF_w	RS	CS	AS
mean	0.62	0.74	0.80	0.79	0.82
median	0.60	0.72	0.81	0.79	0.85
stdev	0.13	0.13	0.13	0.14	0.13

(c)FASR10

	baselines		non-baselines		
	BC_w	BF_w	RS	CS	AS
mean	0.31	0.52	0.68	0.65	0.71
median	0.29	0.50	0.72	0.64	0.75
stdev	0.21	0.21	0.22	0.26	0.23

https://doi.org/10.1371/journal.pcbi.1006813.t002

multiple comparison correction for 6 hypotheses (6 baselines). This correction increases the stringency of the statistical threshold applied on each of the 6 individual tests from $\alpha = 0.05$ to $\alpha = 0.0085$. The CS method also had statistically better ROCAUC performance than all baseline models except BF_l ($p = 0.037$) and BF_s ($p = 0.032$). A complete set of *p*-values from a pairwise comparison of the IBRs is available in S5 Table. The hybrid baseline approaches, which use compound bioactivity profiles to select the most broadly active compounds as informers, performed much better than the chemometric approaches that use chemical features for informer selection.

We also compared strategies using enrichment factor (EF) as an alternative VS metric that, like ROCAUC, reflects retrieval of active compounds (Fig 3). The maximal EF value that could be achieved on a target, however, depends on the active fraction in the set. To address the variation in the extent of the class imbalance across kinase targets (active fractions ranging from 0.01-0.12 in PKIS1) (S2 Fig), we apply the normalized EF metric NEF10. The EF cutoff was also extended from a typical 1% threshold out to 10%, due to the small number of compounds considered ($n = 366$). To simplify comparison with the ROCAUC metric, we scale NEF10 such that a value of 0.5 reflects a random classifier (equivalent to random ranking or no enrichment) and a value of 1.0 represents a perfect classifier, in which the top 10% has been maximally enriched. Over the 224 targets considered in PKIS1, the three bioactivity-based models (RS, CS, and AS) are statistically superior to all of the baseline approaches (all $p < 0.0085$). The AS method had the strongest enrichment for active compounds with a median NEF10 of 0.85 ± 0.13. This was better than the top frequent hitters model, BF_l, which had a median NEF10 of 0.74 ± 0.13 ($p = 5.7E\text{-}14$). The enrichment is even better compared to the chemometric models, the best of which is BC_w, providing a median NEF10 of 0.60 ± 0.13 ($p = 6.7E\text{-}28$).

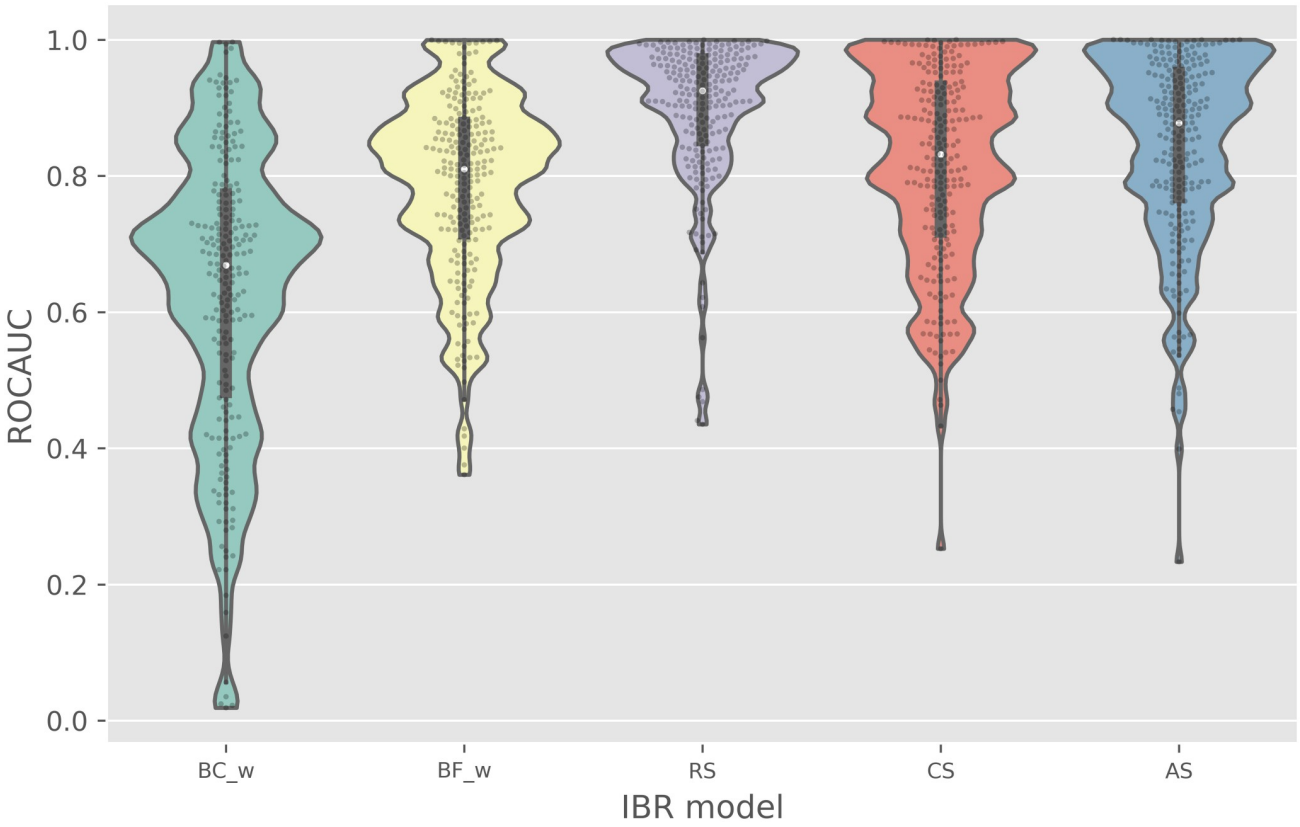

Fig 2. A comparison of models with respect to compound ranking performance as assessed by ROCAUC values. Each model was evaluated on 224 targets through PKIS1 leave-one-target-out validation. ROCAUC of 0.5 indicates a random ranking of compounds on a given target; ROCAUC of 1.0 represents ideal ranking with all active compounds prioritized above the inactives. The individual target evaluations are shown as light grey dots with median and interquartile ranges displayed as a white circle and black bars, respectively.

https://doi.org/10.1371/journal.pcbi.1006813.g002

Another key characteristic of robust virtual screening performance is the recognition of diverse active compound structures rather than retrieval of only a subset of the active chemotypes. Because of the high rate of failure for hits in follow-up hit-to-lead or optimization efforts, we value methods that can retrieve as many active scaffolds as possible, even at some expense to predictive accuracy reflected by ROCAUC and NEF metrics. Across PKIS1 targets we assessed the diversity among the known active chemotypes prioritized by each model by monitoring the Fraction of Active Scaffolds Retrieved among the top ranking 10% of compounds (FASR10) (Fig 4). The bioactivity-based IBR methods outperform the top hybrid and chemocentric baseline models, according to this metric. The median FASR10 for the AS model 0.75 ± 0.23 exceeded the top hybrid model, BF_l (0.52 ± 0.21, p = 2.1E-18), and chemocentric model, BC_w (0.29 ± 0.21, p = 3.7E-32).

Although the IBRs were developed for compound ranking and not necessarily as classifiers, classifier metrics F_1 score (F1) and Matthew's Correlation Coefficient (MCC) were also evaluated for the methods across the PKIS1 targets. The scores/ranks returned by each method were converted to binary classifications using a threshold based on the median active fraction for PKIS1 (5.5%). Over the 224 targets considered in PKIS1, RS, CS, and AS are statistically superior to all of the baseline approaches (all p <0.0085). The full set of metrics evaluations on PKIS1 are provided in S4 Table with the corresponding p-values from pairwise comparisons in S5 Table.

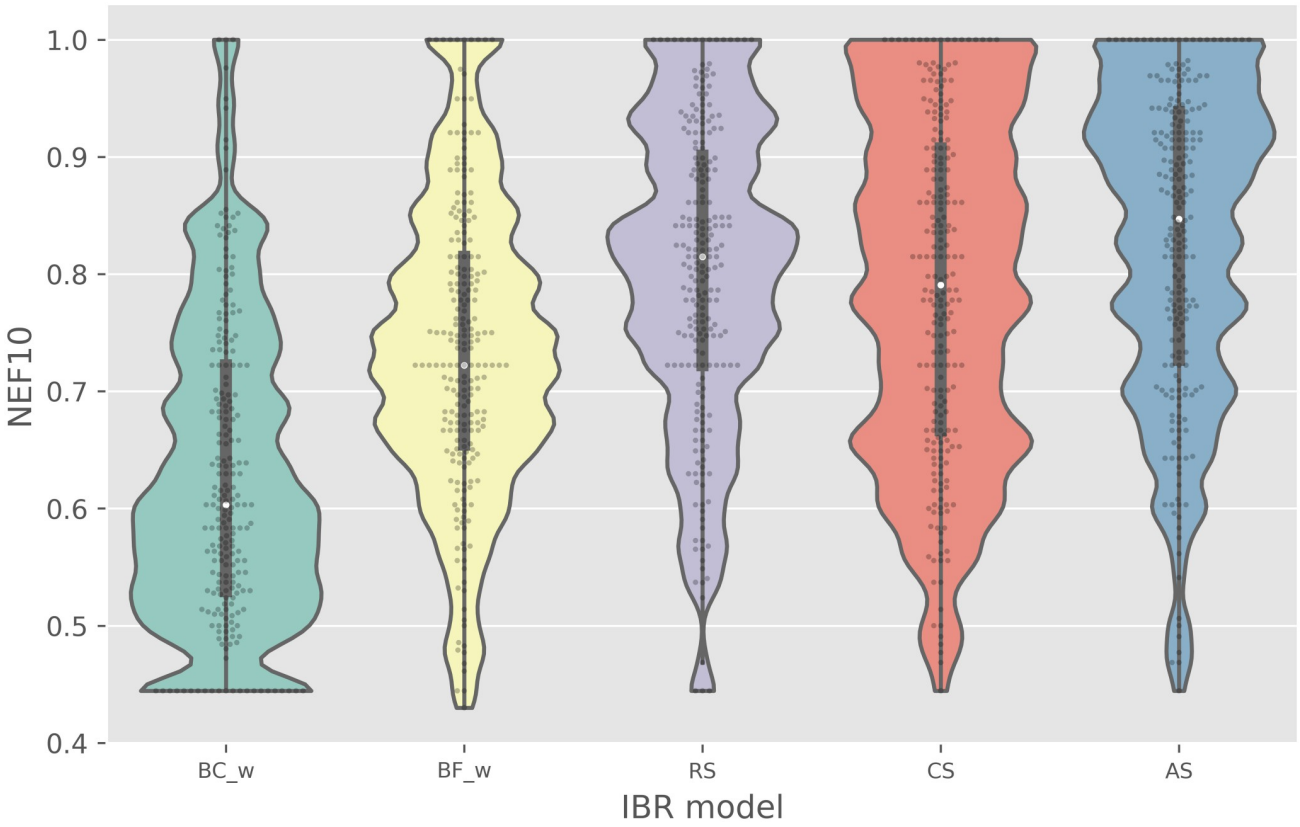

Fig 3. A comparison of models with respect to compound ranking performance as assessed by active enrichment in the top 10% of ranked compounds. Each model was evaluated on 224 targets through PKIS1 leave-one-target-out validation. NEF10 represents the fold-enrichment of actives in top 10% above random that is normalized by dividing by the maximum theoretical fold-enrichment that could be achieved at the 10% threshold for the target of interest.

https://doi.org/10.1371/journal.pcbi.1006813.g003

To assess the robustness of IBR performance, we stratified the PKIS1 targets into four equi-sized subsets and compared IBR methods on all performance metrics separately on each subset. This stratification was based on target hit rate and was obtained by binning targets after ranking by hit rate. S7, S9 and S11 Figs stratify Figs 2–4, and indicate very little effect on performance of the target hit rate. To examine further, we used linear regression to decompose each target-method performance metric into a target effect and a method effect; S8, S10 and S12 Figs plot estimated method effects and multiplicity-adjusted 95% confidence intervals. AS, CS, and RS are all robust to the target hit rate, having quite similar performance in all strata. By contrast, BF_w is relatively sensitive to the target hit rate. Considering that the hit rate of a novel target is unknown prior to testing, marginal features such as in Figs 2–4, reflect relevant operating characteristics of the proposed IBR methods.

Effect of informer set size

All IBR strategies require choosing the number of elements n_A to include in the informer set. Larger n_A allows more information to be gleaned from intermediate screening data, and therefore improved prioritization of non-informer compounds. Marginal improvements in performance as a function of n_A are expected to diminish as n_A increases, because of redundancies in the information acquired as more activity data accrues. Larger n_A also leads to higher assay

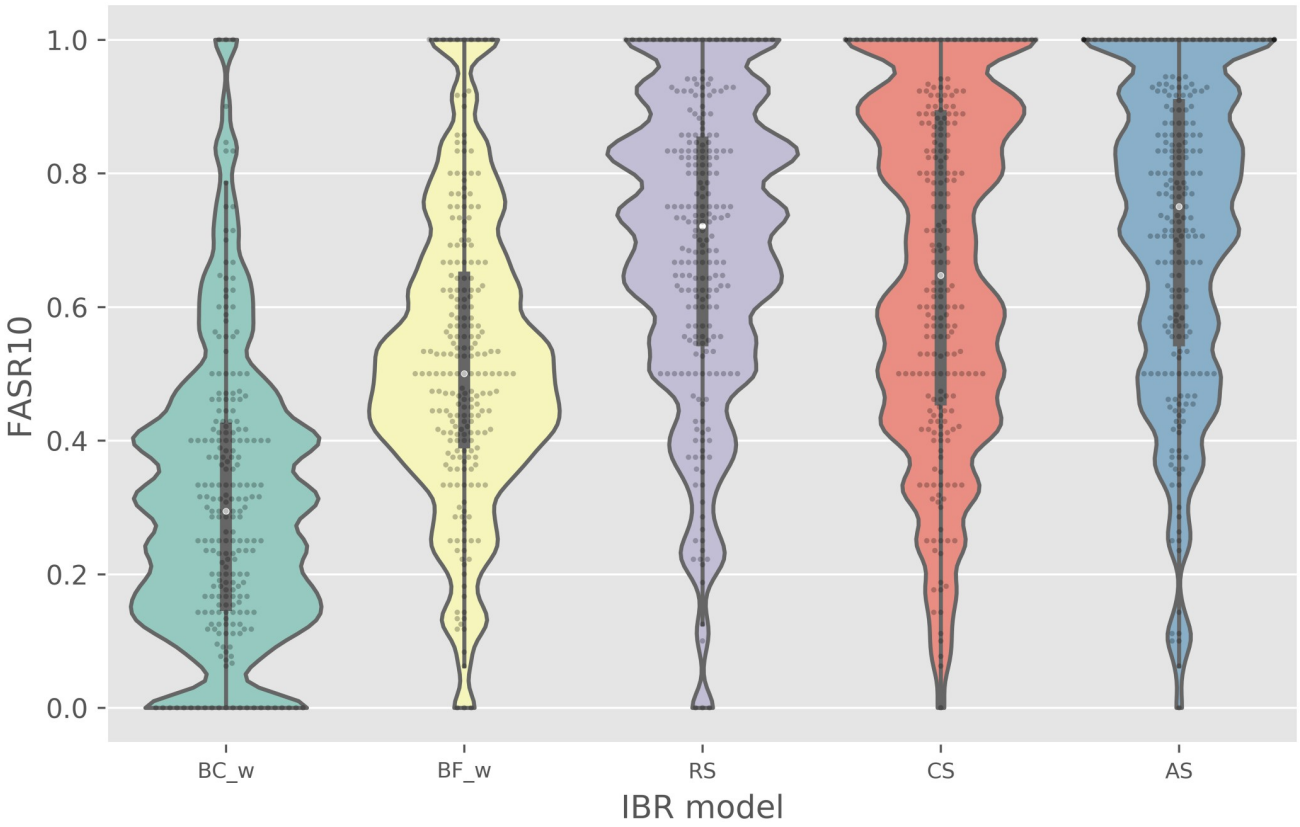

Fig 4. A comparison of models with respect to the structural diversity of the active compounds retrieved. Each model was assessed by FASR10 evaluations on 224 targets through PKIS1 leave-one-target-out validation. The FASR10 metric is the fraction of the total identified active molecule scaffolds, for the target of interest, that were identified in the top 10% of the ranked compounds on that target. Compounds are grouped by their generic (all-carbon skeletons) representations of Bemis-Murcko scaffolds.

https://doi.org/10.1371/journal.pcbi.1006813.g004

costs. The experiments reported above used $n_A = 16$, about 4% of the compounds in the chemogenomics matrix.

To examine the relationship of informer set size to prioritization performance, we applied IBR strategies on a range of informer set sizes. First, we considered AS, our best performing IBR strategy. S4 Fig shows ROCAUC and NEF10 metrics from the LOTO retrospective analysis of PKIS1 for n_A varying from 9 to 28. Performance did not vary greatly over this range. We also tested a wider range of informer set sizes ($n_A = 1$ to 48 compounds) on PKIS1 target predictions using LOTO cross validation, and examined ROCAUC and NEF10 using baseline IBR methods BC_w (S5 Fig) and BF_w (S6 Fig). Over this range, we observe performance degradation with diminishing informer set sizes. These experiments indicate that our preferred value $n_A = 16$ strikes a reasonable balance between size and performance for this particular data set.

Discussion

We set out to establish effective strategies to prioritize compounds for initial testing in iterative high-throughput screens in a drug discovery setting. Our approach is related to the cold-start problem in collaborative filtering (recommender systems) and involves informer-based ranking (IBR) strategies that identify a small subset of highly informative compounds to test in the initial screening round. Data obtained by testing the informers can be used to prioritize

compounds for subsequent screening. As a proof of concept, we focused on kinases so that we could test methods using public kinase chemogenomic data matrices. Among the IBR strategies tested, we found that those leveraging bioactivity data from matrix targets (RS, CS, and AS) provided better initial sampling than baseline strategies that applied chemometric similarity methods (BC) or hybrid approaches (BF). The hybrid approaches used a "frequent hitters" heuristic for informer selection, based on matrix activities, and chemometric similarity for ranking.

We applied our chemogenomic IBR and baseline methods in prospective tests on three microbial kinases: PknB, BGLF4, and ROP18. An initial batch of just 16 informer compounds from each set (roughly 4% of the complete set of compounds) was selected for assays on these new targets. The methods were evaluated with regard to hit prioritization and diversity of active scaffolds prioritized, compared to the results of assays of the PKIS1 and PKIS2 compounds. Results from these prospective tests indicated that IBRs using bioactivity data and hybrid baseline IBRs outperformed baseline IBRs that use purely chemometric data for PknB and BGLF4. The baselines were superior on ROP18 but performed so poorly on PKIS2 compounds for PknB and BGLF4 that they would be risky to apply in practice on a new target.

For a more complete assessment of the IBRs, we performed a retrospective leave-one-target-out validation on the PKIS1 matrix (m = 224 targets by n = 366 compounds) using a batch selection of 16 informer compounds. We observed statistically better hit prioritization and active scaffold retrieval for the purely bioactivity-based IBRs (RS, CS, and AS) than for any of the baseline methods. The successful early hit and active scaffold retrieval in these small kinase datasets suggests that the IBRs could be a valuable approach for prioritizing compounds in larger libraries that cannot be exhaustively screened.

IBRs and related approaches

Chemogenomic assay data have been used through inductive transfer or transfer learning approaches to make successful predictions on compound-target interactions in several contexts [24, 25]. Reker et al. [16] and Cichonska et al. [26] placed chemogenomic predictions into 4 classes: (1) filling in missing elements within a relatively complete chemogenomic matrix (bioactivity imputation), (2) predicting interactions for a target on matrix compounds (virtual screening), (3) predicting interactions for a compound on matrix targets (drug re-purposing or off-target effects), and (4) predicting interactions for non-matrix compounds on a non-matrix target (virtual screening). Wasserman et al. [27] showed that simple kernel approaches using nearest proxy targets could be used to rank compounds effectively for a query target (class 2), as long as it was possible to identify proxy targets closely related to the query target. For kinase targets, Cichonska et al. [26] explored a wide range of ligand and target kernels to address class-1 and class-3 problems. For focused target sets (kinases and GPCRs), Janssen et al. [28] recently applied nearest-neighbor approaches to ligand and targets mapped on t-SNE projections to address class-2 and class-3 problems.

The methods we report differ from prior chemogenomic methods for addressing the class-2 problem by involving strategic but limited data acquisition on the query target. Determination of the responses of targets to key informer compounds shifts a relatively difficult class-2 problem into the more tractable class-1 problem of imputation. Unlike chemogenomic kernel-based approaches [26, 27], we did not use target features, focusing instead on target-agnostic strategies for compound ranking that could be used in the future for cell-based or phenotypic assays. Our focus on limited, strategic data acquisition on the target of interest frames the problem in a more practical context akin to compound prioritization in early, low-data stages of an iterative screening effort [14, 18, 29, 30]. Lack of active compound instances can stall

implementation of supervised models for compound selections [15]. Our bioactivity-based IBR methods overlap hit expansion methods using chemogenomic data, as applied by groups at Novartis for guiding molecule selection in iterative screening [18, 29]. In agreement with their findings, IBR methods that use compound bioactivity profiles, rather than chemical features, provided broader active scaffold retrieval [29]. Previous implementations of HTSFP, however, define compounds by normalized bioactivity vectors from an independent reference assay set, whereas our IBRs use compound bioactivity profiles derived directly from the available chemogenomic matrix. We tested targets only from the same target class, namely, protein kinases. The IBR-based informer sets could be applied in the same way that Paricharak et al. used their Mechanism-of-Action Box (MoABox) of probe compounds for testing in "iteration zero" of their iterative screening procedure [29].

Practical implications

The IBR strategies described here could enable iterative screens either on orphan members of a target class or on targets on which very few compounds have been tested. Data returned on each screening iteration would then be used as new training instances to refine the model, potentially in an active-learning framework that also considers relevance of training instances for subsequent compound selection. To promote efficiency of an iterative approach, initial compound batches are often limited in size, with compounds are often being selected at random or to achieve chemical diversity. Initial screens chosen in this way are likely to return few active compounds, thus stalling effective implementation of a supervised activity prediction model. The IBR strategy reported here can be deployed for compound prioritization in early rounds of batch selection; the informer set could be tested to obtain preliminary compound rankings in the low-data phase of iterative screens. Due to class imbalance being skewed towards inactive compounds in drug discovery tasks, IBR methods could enable rapid identification of relatively rare but important active instances necessary for training the activity-prediction model until it can score compounds accurately for prioritization.

Moreover, the bioactivity-based IBR methods exhibited diverse active-scaffold recognition properties, yielding positive training instances with greater structural diversity for supervised compound prioritization models. The FASR10 results indicate that bioactivity-based IBR approaches generalize better over different compound structures than chemometric IBRs, so they should exhibit a greater tendency to scaffold hopping [29, 31, 32]. In contrast, all of the baseline IBR methods use Morgan fingerprint-derived distances to active informers, thus confining their perspective to those active regions of chemical space identified with the informer set. Different chemotypes, however, can exhibit strong activity on the same target. Plots of PKIS1 compounds projected into their three major principal components of chemical feature space (Morgan fingerprints) frequently show active regions that are non-adjacent (S3 Fig). While active compounds tend to cluster in specific regions of chemical space, many targets elicit multiple, sometimes distantly separated regions of active chemical space.

Future directions

There are several potential uses for IBRs in drug discovery. This work demonstrates the possibility of effective prediction of activities for new targets within the same target class (kinases) from an extensive chemogenomics data matrix representing many targets within that class. A future direction of research is to quantify the amount of chemogenomic data needed to enable robust prediction within the same target class. It appears that low-rank structure in the chemogenomic matrix used in the IBR methods helps to enable reliable predictions of a target's compound preferences. Statistical models that faithfully represent variation and dependence in

bioactivity data also could be leveraged to guide the development of alternative IBR strategies beyond RS, CS, and AS.

Of greater interest is the development of a more general informer set from a broader collection of chemogenomic data. To investigate the generalizability of the methods, we plan to apply them to a wider range of novel targets (or held-out targets) using an expanded chemogenomic data set with broader target and compound coverage. We do not know how well IBRs will perform on new targets that are unrelated to those within the matrix. We are encouraged by the prospective predictive performance on query kinases (PknB, BGLF4, and ROP18) that are dissimilar from kinases in the chemogenomic data but note that these targets are still similar structure and chemical function as protein kinases. More comprehensive data matrices tend to be incomplete, with many missing data values, but they should be useful in testing whether these methods are effective in extended pharmacological spaces. The size of the informer set may well have a dramatic impact on overall performance.

It may be possible to use IBR methods for prioritizing non-matrix compounds on a new target (a class-4 problem). Chemogenomic matrices enable pharmacological mapping of a given new target (query) to matrix targets that exhibit similar bioactivity profiles (proxy targets). Associations between query targets and proxy targets can be made on the basis of full-compound bioactivity profile in the matrix, or potentially just informer assay results. Given that certain proxy targets are likely to be more extensively screened (tested with compounds outside the matrix set), it might be possible to use non-matrix screening data on proxy targets to infer activities for additional compounds and thus prioritize them for testing on some query target.

Materials and methods

PKIS data sets

Most of the IBR strategies developed here leverage chemogenomics data matrices for activity predictions on compounds against selected kinase targets. The matrices were derived from two public human kinase chemogenomics data sets PKIS (PKIS1) [20, 33] and PKIS2 [21]. Prior to development and testing of methods, these sets were processed as described below. (Links to our processed PKIS datasets are provided below).

PKIS1. The original PKIS data set (PKIS1) was downloaded from https://www.ebi.ac.uk/chembldb/extra/PKIS/PKIS_screening_data.csv Each row in this data set contains an assay result on a specific compound. Each row lists several identifiers for each compound and the target, assay conditions, and the assay read-out (percent inhibition). For nearly every compound, kinase activity was tested independently at 0.1 μM and 1.0 μM concentrations. For this work, only the inhibition values obtained at 1.0 μM were used, in order to match the PKIS2 concentrations. PKIS1 contains 366 unique compounds with unique SMILES and ChEMBL IDs that were tested on 200 unique parent kinases having unique target ChEMBL IDs. When we include mutants/variants of the parent isoforms, there is a total of 224 targets with unique ChEMBL ASSAY IDs. Our processed PKIS1 data was therefore arranged as a matrix of 224 kinase targets by 366 compounds.

PKIS2 The original PKIS2 was downloaded from https://doi.org/10.1371/journal.pone.0181585.s004. This set comprises 641 unique compound SMILES and 406 target columns. However, only of these 415 compounds were available to us from the original set for testing. We included only these compounds from the PKIS2 data set, so our bioactivity matrix has 406 targets by 415 compounds. PKIS2 activity values represent percent inhibition values observed at inhibitor concentrations of 1 μM.

Bioactivity-based prediction methodology

Setup. We let $X = \{x_{i,j}\}_{i=1}^{m}, {}_{j=1}^{n}$ denote the bioactivity inhibition matrix that is available initially, where m denotes the number of kinase targets and n denotes the number of compounds. We use $\boldsymbol{x}_{i\cdot}$ for the vector of bioactivity results on target i, and $\boldsymbol{x}_{\cdot j}$ for column entries of this matrix (that is, bioactivity results for compound j). Let $I = \{1, 2, \cdots, m\}$ denote the targets associated with rows of the data matrix X, and $J = \{1, 2, \cdots, n\}$ denotes the set of available compounds.

For some methods, the kinase inhibition matrix X is reduced to a binary matrix $Z = \{z_{i,j}\}_{i=1}^{m}{}_{j=1}^{n}$, which captures empirical assessments of whether target i is inhibited (or not) by compound j. We use a target-wise threshold criterion Eq (1), based on the sample mean and sample standard deviation of each row $\boldsymbol{x}_{i\cdot}$, as follows:

$$z_{i,j} = \begin{cases} 1, & \text{if } x_{ij} \geq \text{mean}[\boldsymbol{x}_{i\cdot}] + 2 \times \text{s.d.}[\boldsymbol{x}_{i\cdot}], \\ 0 & \text{if } x_{ij} < \text{mean}[\boldsymbol{x}_{i\cdot}] + 2 \times \text{s.d.}[\boldsymbol{x}_{i\cdot}]. \end{cases} \tag{1}$$

Our ultimate task is prediction from X of binary activities: $z_{i^*,j}$, on a new target $i^* \notin I$ for available compounds, $j \in J$. Our approach is first to identify from X a small *informer set* of compounds, $A \subset J$, on which bioactivity data $x_{i^*,j}$ will be measured against target i^*. The data obtained from this experiment with the informer set, denoted by $\boldsymbol{x}^* = \{x_{i^*,j}, j \in A\}$, will be used to identify other compounds in the full compound set J that inhibit new target i^*, in the sense that $z_{i^*,j} = 1$. Machine-learning and statistical tools are used to design and study this approach, but except through general parallels with adaptive experimental design, the selection of an informer set is neither a supervised nor an unsupervised machine-learning task. We describe three novel heuristic methods that have favorable empirical characteristics: regression selection (RS), coding selection (CS), and adaptive selection (AS).

The informer-based ranking (IBR) methods that we propose entail partitions of the target set I, also referred to as a set of clusters, sometimes denoted by $S = \{S_1, S_2, \cdots, S_K\}$. Methods differ as to how any candidate partition S is evaluated or acted upon. A partition S induces a labeling of targets $i \in I$, denoted $y = (y_i)$, where $y_i = k$ if (and only if) $i \in S_k$. In all methods, the new target i^* becomes associated with one of the clusters by virtue of similarity of bioactivity profiles with other targets.

Regression selection (RS). The relationship between the training target space I and a new target (point) needs to be established in order to predict active compounds on novel kinase targets. The informer set serves to locate the new target in the training space. Unsupervised clustering is used to partition the target space; then the informer set is chosen from compounds that are predictive of cluster labels in a coupled, supervised analysis.

Informer set. The informer set is identified using clustering, regression, and feature selection. First, we classify the target space—the row space $\boldsymbol{x}_{i\cdot}$—into clusters such that all targets within the same cluster exhibit a similar response to the compounds. For this task we considered `k-means`, which tries to minimize the sum of the within-cluster distances from each cluster centroid. Formally, given a parameter K as the number of clusters, and m data vectors $\boldsymbol{x}_{i\cdot}, \ldots, \boldsymbol{x}_{m\cdot}$, it aims to solve

$$\min_{S} \sum_{k=1}^{K} \sum_{p \in S_k} \left\| \boldsymbol{x}_{p\cdot} - \frac{\sum_{q \in S_k} \boldsymbol{x}_{q\cdot}}{|S_k|} \right\|^2, \tag{2}$$

where $S = \{S_1, \ldots, S_K\}$ forms a partition of $I = \{1, \cdots, m\}$, and $|V|$ denotes the cardinality of the

set V. For robustness, we scale each column of the bioactivity data X linearly so that its entries lie in the range [0, 1] prior to clustering analysis.

There are two difficulties in using `k-means`. The first is that its iterative process relies on random initialization, hence the results generally differ on each run. Secondly, we do not know in advance how to specify the number clusters K. To deal with the first problem, we use the `kmeans++` initialization procedure [34]. `kmeans++` guarantees that the expected final objective value is no more than $O(\log K)$ times larger than the optimal. To further improve robustness, we repeat the `kmeans++` procedure 100 times and choose the outcome that has the lowest objective value in Eq 2. For the second issue of selecting K, we find the value that achieves the best performance in a five-fold cross-validation procedure.

We note that other techniques for clustering are available, for example, hierarchical agglomerative clustering. None of these techniques has a strong guarantee of finding a global minimizer of Eq 2, and some are deterministic rather than stochastic, yielding only one candidate solution. Our approach based on `k-means` is preferable because it can be run multiple times, cheaply, generating numerous candidate solutions, of which the one achieving the minimum value of Eq 2 can be chosen. In practice, because many instances of our procedure converge to the same conformation with the apparent global minimum value of Eq 2, we are confident that it finds the global solution.

Clusters serve to label the targets, as noted above. Namely, we set $y_i = k$ if $i \in S_k$ in Eq 2. Next, multinomial logistic regression with a penalty term is applied to train a label classifier. In this approach, training data has the form $\{(\boldsymbol{x}_{i\cdot}, y_i)\}$, over an appropriate set of targets i. The multi-class classifier is trained by fitting the multinomial logistic model. That is, we seek a set of coefficients, $\boldsymbol{\omega} = \{\omega_{10}, \cdots, \omega_{K0}, \boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_K\}$, by minimizing the objective function

$$-\log\left(\prod_i \frac{\exp\left(\omega_{y_i 0} + \boldsymbol{w}_{y_i}^T \boldsymbol{x}_{i\cdot}\right)}{\sum_{k=1}^{K} \exp\left(\omega_{k0} + \boldsymbol{\omega}_k^T \boldsymbol{x}_{i\cdot}\right)}\right) + \lambda R(\boldsymbol{\omega}), \tag{3}$$

where the first term is the (negative) log-likelihood from the multinomial logistic model, $\lambda \geq 0$ is a tuning parameter, and $R(\boldsymbol{\omega})$ is a penalty function. The coefficients $\boldsymbol{\omega}_k$, one for each cluster, are vectors whose length equals the number of compounds, each of whose elements $(\boldsymbol{\omega}_k)_j$ represents the weight that is applied to the activity measurement for compound j in predicting membership of cluster k. An appropriately chosen penalty yields sparse solutions in which coefficients of compounds that are only weakly predictive of the target cluster are set to zero. Since the whole coefficient vector $\boldsymbol{\omega}_{\cdot j} := ((\boldsymbol{\omega}_1)_j, (\boldsymbol{\omega}_2)_j, \cdots, (\boldsymbol{\omega}_K)_j)$ captures the influence of compound j on all targets, variable selection is achieved only if these K coefficients are shrunk simultaneously to zero. Therefore, we use the group LASSO norm [35] corresponding to the penalty function:

$$R(\boldsymbol{\omega}) = \sum_{j=1}^{n} \|\boldsymbol{\omega}_{\cdot j}\| = \sum_{j=1}^{n} \left((\boldsymbol{\omega}_1)_j^2 + \cdots + (\boldsymbol{\omega}_K)_j^2\right)^{1/2}, \tag{4}$$

where $\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_2 = \sqrt{\boldsymbol{u}^T\boldsymbol{u}}$ denotes the L_2-norm.

We combine this regularized model with the greedy heuristic proposed in [36], which was shown to outperform the model obtained by directly solving Eq 3 by choosing relevant features greedily, one at a time. This algorithm starts by setting $A = \emptyset$, then solves Eq 3, and then selects the feature vector $\boldsymbol{\omega}_{\cdot j}$ with the largest Euclidean norm, from among those features still represented in the regularization term R. It adds j to the informer set A, then re-solves Eq 3 with all feature vectors $\boldsymbol{\omega}_{\cdot j}$ for $j \in A$ excluded from R. This process is repeated until either we have selected enough features (denoted by n_A) or else the remaining $\boldsymbol{\omega}_{\cdot j}$ included in the norm

calculation are all zero. After selecting the features in this fashion, we retrain a model by solving Eq 5 again using *only the selected features* and omitting the regularization term R altogether, that is,

$$\min_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_k} \; -\log\left(\prod_{i=1}^{m}\frac{\exp\left(\omega_{y_i 0}+\boldsymbol{w}_{y_i}^T\boldsymbol{x}_{i\cdot}\right)}{\sum_{k=1}^{K}\exp\left(\omega_{k0}+\boldsymbol{w}_k^T\boldsymbol{x}_{i\cdot}\right)}\right), \quad \text{such that } \boldsymbol{\omega}_{\cdot j}=0 \text{ for } j\notin A. \tag{5}$$

Compound ranking. Given any new target and its bioactivity with compounds from the informer set, we use the trained logistic regression model to predict the cluster label. With y_{i^*} denoting the to-be-predicted cluster label of the new target i^* and $\boldsymbol{x}^*$ the informer set compound activities for this target (i.e., the intermediate data), the multi-class logistic model asserts:

$$\Pr(y_{i^*}=k|\boldsymbol{x}^*)=\frac{\exp\left(\boldsymbol{w}_k^T\boldsymbol{x}^*\right)}{\sum_{l=1}^{K}\exp\left(\boldsymbol{w}_l^T\boldsymbol{x}^*\right)}. \tag{6}$$

Using this probability and the centroids, we predict the whole activity vector of the new target with the vector of expected values:

$$\hat{\mu}_j \coloneqq \sum_{k=1}^{K}\Pr(y_{i^*}=k|\boldsymbol{x}^*)\frac{\sum_{i\in S_k}x_{ij}}{|S_k|} \tag{7}$$

for compound j.

Parameter estimation. Given the whole active prediction procedure, we are able to conduct parameter selection using cross-validation. We conduct five-fold cross-validation to pick the best value of K for each evaluation metric and then retrain on the whole data matrix using the selected K to generate the final model for predicting the test data. Pseudo-code for the entire regression selection process is provided in Algorithm 1 in the S1 Text. We fix $\lambda = 10^{-6}$ in Eq 3 in our implementation.

Coding selection (CS). Coding selection works directly on the binary bioactivity data $Z = \{z_{i,j}\}_{i\in I, j\in J}$ rather than the quantitative inhibition measurements. The idea is to construct a single objective function to score potential informer compounds for how well they predict target activity the non-informer set. For a potential informer set $A \subset J$, CS considers that distinct rows of the sub-matrix $Z_A = \{z_{i,j}\}_{i\in I, j\in A}$ constitute a kind of encoding of the kinase target space. Specifically, row i of Z_A, which corresponds to kinase target i, is a length $n_A = 16$ vector of zeros and ones. Among the $2^{16} = 65536$ possible such vectors, only a relatively few distinct ones, numbering $L_A \leq n$, manifest themselves as rows of the sub-matrix in a given example. We call these distinct vectors *code words*, and denote them $q_1, q_2, \cdots, q_{L_A}$. For some $K \leq L_A$, we introduce a partition $\pi = \{b_k\}$ of these code words, where each block b_k holds a set of code words, and where π has K disjoint blocks. Together, the informer set A and the partition π induce a partition $S = \{S_1, S_2, \cdots, S_K\}$ of the targets I by the rule that $i \in S_k$ if (and only if) row i of the sub-matrix equals some code word $q_l \in b_k$. We emphasize that given candidates A and π, the target-space partition S is obtained using only information in the binary data sub-matrix Z_A.

To provide some intuition for the coding construction, let's look ahead to when intermediate data $\boldsymbol{x}^*$ are obtained in experiments with informer set compounds $j \in A$ on new target i^*. These inhibition measurements may also be binarized to produce bioactivity calls $z^* = \{z_{i^*,j}, j \in A\}$. If z^* exactly matches one of the code words q_l, then any targets in I having this same code word are natural comparators for i^*. Their bioactivity profiles on the non-informer compounds may be the basis for a useful secondary prediction. In fact we may not

have an exact match of the new code word, and there may be distinct code word profiles on the informer compounds that yield similar non-informer profiles. Therefore, we propose the following objective function to measure properties of the potential informer set A and the code-word partition π that are conducive to high-accuracy prediction on non-informer compounds:

$$f_{K,\lambda}(A,\pi) = \sum_{k=1}^{K}\left(\sum_{i,i'\in S_k}\left\{1 - \frac{\sum_{j\in A^c} z_{ij}z_{i'j}}{\sum_{j\in A^c} z_{ij} \vee z_{i'j}}\right\}\right) - \lambda L_A \tag{8}$$

The inner summation in Eq 8, which is over pairs of targets within cluster S_k, accumulates pairwise differences between targets, as measured on the non-informer compounds, and using the *asymmetric binary* distance: among non-informer compounds that are active against either target, what fraction are not active against both? The outer sum is over clusters (partition blocks) of targets. The objective function value is low, therefore, if clusters induced by A and π and informer data are internally homogeneous on the non-informer data. For tuning parameter $\lambda > 0$, the penalty term λL_A encourages informer sets that have many code words, in order to reduce the rate of extrapolation from intermediate data. The proposed scoring function Eq (8) is essentially non-parametric, allowing potentially complex relationships to exist between bioactivities of informer and non-informer compounds. This modeling flexibility comes at a cost, however, in that it is a combinatorial optimization task to identify the best A and π settings for any fixed K and λ.

Initially we sought to solve $\arg\min f_{K,\lambda}(A, \pi)$ approximately by Monte Carlo search. Fixing parameters K and λ, we randomly sample $(A_1, \pi_1), (A_2, \pi_2), \cdots, (A_B, \pi_B)$ for a large number of trials B, such as 10^6 or 10^7. Each A_b is a random subset of size $n_A = 16$ taken from the full set of n compounds; then π_b is a random partition of the code words from Z_{A_b}. In numerical experiments, we found that marginally stabilizing compound scores is more effective than taking the informer set $\hat{A}$ to be the sampled set A_b having the lowest objective value Eq (8). Specifically, we score every compound $j \in J$ by

$$f_j = \sum_{K\in\mathcal{K}} \frac{1}{B}\sum_{b=1}^{B} 1(j \in A_b) f_{K,\lambda}(A_b, \pi_b) \tag{9}$$

where $\mathcal{K}$ is a set of entertained cluster numbers. We used $\mathcal{K} = \{2, 3, \cdots, 40\}$, and fixed $\lambda = 5$ based on preliminary experimentation. The computed informer set $\hat{A}$ contains the n_A best (lowest) scoring compounds by this score.

Compound ranking. To proceed with ranking compounds, we require a code word z^* derived from the intermediate data x^* obtained on the new target i^*. A threshold level, such as used to binarize the original data Eq (1), may not be available. Instead we revert to the inhibition data on the informer compounds, say $X_{\hat{A}}$ (an $m \times n_A$ sub-matrix of X), and we keep track of all the rows of $X_{\hat{A}}$ associated with each code word in the computed informer set. We compute a centroid for code-word q_l, say, c_l, by averaging the rows of $X_{\hat{A}}$ associated with q_l. Then, the code-word centroid that is closest (in Euclidean distance) to the new data x^* is the derived code word z^* for target i^*.

Having our new target i^* provide code word z^* on the basis of intermediate data x^*, we next require a prediction of non-informer compounds that may also inhibit i^*. We score $j \in \hat{A}^c$ by

their activity rates among the n^* targets with the same code word as i^*:

$$a_j = \sum_{i \in I} z_{i,j} 1[z^* \text{ is the code word of } i\,], \quad n^* = \sum_{i \in I} 1[z^* \text{ is the code word of } i\,]. \tag{10}$$

Our prediction of the active non-informer compounds is $\mathcal{L} = \{j \in \hat{A}^c : a_j \geq \kappa\}$ for some threshold κ. We set κ with an appeal to false-discovery-rate (FDR) control, recognizing that the Bernoulli trial $z_{i^*,j}$ may be regarded as having success probability estimated by a_j/n^*. Then a crude estimate of FDR of $\mathcal{L}$ is

$$1 - \frac{\sum_j (a_j/n^*) 1[a_j \geq \kappa]}{\sum_j 1[a_j \geq \kappa]} \tag{11}$$

Similarly, we could estimate under the Bernoulli model the expected number of active non-informer compounds, $\sum_{j \in A^c} (a_j/n*)$, which may guide our choice of κ. Pseudo-code for the entire coding selection method is provided in Algorithm 2, S1 Text.

Adaptive selection (AS). The AS approach first identifies a base informer set of size $n_0 < n_A$ compounds by a minor variation of the regression selection (RS) approach. We use $n_0 = 8$ and $n_A = 16$. This step establishes both a clustering of the target space and the identity of n_0 compounds that are predictive of the cluster labels. Next, AS adaptively grows the informer set, one compound at a time, so as to identify compounds that are predictive of non-informer bioactivity.

To identify the base informer set A_o, the target space I is clustered using `k-means`, which aims to solve Eq 2. With $L_K = \min_S \sum_{k=1}^{K} \sum_{p \in S_k} \left| \boldsymbol{x}_{p\cdot} - \frac{\sum_{q \in S_k} \boldsymbol{x}_{q\cdot}}{|S_k|} \right|^2$, the number of clusters K is determined by

$$K = \arg\min_k \left\{ \left| 1 - \frac{L_{k+1}}{L_k} \right| \leq \epsilon \right\}. \tag{12}$$

ϵ is a small value. In our calculations, $\epsilon = 0.02$. Similar to RS, the clustered target space $\{S_k\}$ then serves as the response variable in a penalized multinomial regression to select the first n_0 compounds of the informer set. Our specific implementation uses group LASSO as deployed in the `glmnet` R package for the multinomial response [37]. The regularization penalty is chosen so that precisely n_0 compounds enter the predictive model.

For the remaining $n_A - n_0$ informer compounds, we augment the current set one compound at a time. Letting A_c denote the current set, the next added compound solves:

$$j^* = \arg\min_{j \notin A_c} \sum_{k \notin A_c \cup \{j\}} \| \boldsymbol{x}_{\cdot k} - \boldsymbol{c}_n \|_2, \tag{13}$$

where $\boldsymbol{x}_{\cdot k}$ is the column vector in the inhibition matrix $\{x_{i,j}\}_{i \in I, j \in J}$ for compound k; $\boldsymbol{c}_n = \frac{1}{|A_c \cup \{j\}|} \sum_{k \in A_c \cup \{j\}} \boldsymbol{x}_{\cdot k}$ denotes the centroid of the current informer set. Eq 13 finds the compound that minimizes the distance between informers and non-informers; the informer set is updated $A_c \leftarrow A_c \cup \{j^*\}$. The final informer set is generated by iterating this process until there are n_A compounds in A_c.

For compound ranking, AS uses the same approach as CS Eq (11) after the code word z^* is acquired on generated informer set. Pseudo-code for AS is provided in Algorithm 3 of S1 Text.

Baseline models (B). As practical baseline approaches against which to compare our bioactivity-guided experimental design strategies (RS, CS, and AS), we applied two different informer selection methods: one based on compound structural diversity and the other

leveraging the most frequent hitter (nonselective) kinase inhibitors as observed from the compound bioactivity matrix. Then, based on data returned from these informer selections, we applied three different chemical feature-based compound ranking methods, yielding a total of six strategies.

Baseline informer set selection. Baseline informer compounds were selected from each data matrix by one of two different methods:

- *Chemometric* selection (*BC*)—Compounds are grouped by scikit-learn's hierarchical agglomerative clustering procedure (n_A = 16 clusters, average linkage) using a Jaccard distance matrix computed from RDKit-derived Morgan chemical fingerprints (radius = 2, 1024-bits) as features [38–40]. The 16 cluster medoids are taken as the informers.
- *Frequent Hitters* selection (*BF*)—Matrix compounds *j* are ranked in descending order by the number of targets on which each is labeled active, i.e., by $f_j = \sum_{i \in I} z_{i,j}$, where $z_{i,j}$ indicates activity in the input bioactivity data (1). In other words, the informer set contains the 16 most broadly active compounds.

Note that in the cross validation study, the informer set needs to be recomputed each time a different target is left out.

Baseline compound ranking. After data were returned on the informer set, the remaining non-informer compounds were then ranked by three chemometric "hit expansion" methods:

- *Simple Expansion (s)*—ranks each non-informer by its distance to the nearest active informer compound as measured by Jaccard distance between Morgan fingerprints.
- *Loop Expansion (l)*—loops through active informer compounds in order of descending activity and prioritizes the nearest unranked non-informer compound to the current active informer based on fingerprint distance. The loop continues until all non-informers have been ranked.
- *Weighted Expansion (w)*—ranks each non-informer compound by Euclidean inner product of a target's informer activity vector (16 normalized activities) and a compound's vector of Jaccard similarities to those 16 informers. This scalar represents the "activity-weighted" similarity of each compound to the informer set. Compounds are prioritized in order of ascending values. Therefore, for target i^*, we have a single activity vector $\boldsymbol{x}^*$ comprising the 16 informer activities (normalized to [0, 1]):

$$\boldsymbol{x}^* = [x_1, x_2, x_3, ..., x_{16}] \tag{14}$$

Each noninformer compound, *j*, has a similarity vector $\boldsymbol{v}$ representing the compound's similarity to each of the informers tested on target i^*:

$$\boldsymbol{v}_j = [v_1, v_2, v_3, ..., v_{16}] \tag{15}$$

The weighted expansion score, *w*, for compound *j* on target i^* is then the Euclidean Inner product:

$$w_j = \boldsymbol{v}_j^T \boldsymbol{x}^* \tag{16}$$

In the simple and loop expansion ranking methods, a binary label is required to designate "active" and "inactive" informers. One issue that arises from this is that a compound's activity label depends on a target's compound activity distribution (μ+2σ threshold), which is unknown prior to experimental screening of the compound set. Another is that these

expansion methods cannot proceed if none of the informers are identified as active. To address the former issue, we predict the activity threshold used for assigning a compound's binary activity label on a given target using data returned on the 16 informer compounds. From these 16 informer activities, a threshold is inferred from the available compound activity distributions on other targets in the matrix. This threshold for each target is the κ parameter described above in the Coding Selection method. To address the latter issue, where no active informer compound is identified, the simple and loop expansion ranking methods treat the highest activity informer compound as an active center for expansion.

Biological assays on novel microbial kinase targets

Mycobacterium tuberculosis PknB. Recombinant bacterial kinase (PknB) and bacterial substrate (GarA) were purified from *E. coli* following published procedures [22]. The kinase inhibition assay was done using the Kinase Glo(R) kit from Promega similar to published procedures. PknB was added to plated kinase inhibitor libraries (the available compounds from PKIS 1 and 2) and incubated at room temperature for 10 minutes, after which ATP and GarA (protein substrate) were added. The final concentrations were: PknB 0.25 μM, GarA 40 μM, ATP 100 μM, inhibitors 2 μM, DMSO 1% in a final volume of 5 μL. The kinase reaction proceeded at room temperature for 30 minutes and quenched by the addition of 5 μL of Kinase Glo(R) reagent. The plate was allowed to develop for 10 minutes and luminescence was detected on a BMG PheraStar multiplate reader. Luminescence was converted to μmol/minute of ATP consumed using a standard curve of ATP from 100 to 0 μM. A negative control (no inhibitor) was used to determine percent activity. A positive control (GSK690693) was used to ensure a baseline and compare plate-to-plate variation. Data were analyzed using CDD Vault (Collaborative Drug Discovery, Inc.) to determine plate $Z' > 0.5$ and report percent inhibition for each compound.

Epstein-Barr virus BGLF4. Viral kinase BGLF4 was provided by the laboratory of Professsor Yongna Xing. BGLF4 was expressed with an N-terminal His_8-MBP-dual-tag in insect cells, and purified over Ni^{2+}-NTA resin (Qiagen) and then Maltose resin (Qiagen), followed by ion exchange chromatography (Source 15Q, GE Healthcare) and gel filtration chromatography (Superdex 200, GE Healthcare) to more than 95% homogeneity. The purified BGLF4 was then used for kinase inhibition assays using the C-terminal fragment peptide of retinoblastoma protein (RB) as substrate (Millipore Sigma cta# 12-439). The remaining assay parameters were the same as those applied for PknB except for the following changes. The final concentrations in the reaction medium were: BGLF4 0.004 μg/μL, RB 0.04 μg/μL, ATP 500 μM, inhibitors 3 μM, DMSO 0.3% in a final volume of 5 μL. As a positive control, K252a (5 μM) was used. The reaction proceeded at room temperature for 20 minutes and was then quenched by the addition of 5 μL of Kinase Glo(R) reagent. ADP depletion proceeded for 40 minutes, followed by addition of 10 μL of kinase detection reagent. The reactions were incubated for 1 hour prior to luminescence detection.

Toxoplasma gondii ROP18. Inhibition data for the PKIS compounds on the *Toxoplasma gondii* kinase ROP18 was provided by the University of North Carolina Structural Genomics Consortium and Professor L. David Sibley at Washington University in St. Louis. Their assay measured phosphorylation of a substrate peptide by purified ROP18 using microfluidic capillary electrophoresis [23].

Model metrics and evaluation procedure

Metrics. To evaluate model performance, we applied three different virtual screening metrics, ROCAUC, NEF10, and FASR10. Standard classification metrics, F_1 score (F1) and Matthew's

Correlation Coefficent (MCC), were applied as well. ROCAUC and NEF10 measure the extent to which a model prioritizes the active compounds in its ranking. ROCAUC is a standard metric in virtual screening [41] and applied generally in machine learning to evaluate classifiers. Enrichment Factor (EF) (Eq 17) is another commonly used metric for assessing virtual screening performance. EF reflects the fold increase in active compounds over that expected from random compound selection, for a subset of a compound library taken from some top ranking portion of a prioritized compound list.

$$\mathrm{EF10}_i = \frac{\sum_{j \in B} z_{i,j}}{|B|} \Big/ \frac{\sum_{j=1}^{n} z_{i,j}}{n}, \tag{17}$$

where B is the set of compounds among the top 10% of those ranked by a method applied to target i, and $z_{i,j}$ is as in (1).

However, the number of active compounds for each left-out target i varies from target to target (S1 Fig). We apply a scaling scheme on EF at the top 10% (Eq 18), which enables better comparisons across targets exhibiting significant differences in active:inactive ratios.

$$\mathrm{NEF10}_i = \frac{1 + \dfrac{\mathrm{EF10}_i - \mathrm{EFbase}}{\mathrm{EF10max}_i - \mathrm{EFbase}}}{2}, \tag{18}$$

where EFbase is 1, which corresponds to random guessing; $\mathrm{EF10max}_i$ is the maximum theoretical $\mathrm{EF10}_i$, which means all actives are ranked at the top and depends on the number of actives for each target. Our NEF metric returns a value between 0.5 and 1, where a $\mathrm{NEF10}_i$ larger than 0.5 shows better ranking performance than random guessing–similar to ROCAUC. We selected the 10% threshold with consideration of the sizes of our informer ($n_A = 16$) and full compound sets ($n = 366$ and $n = 405$). This threshold includes the 16 informers and 21 noninformer compounds in our PKIS1 evaluation.

For the ROCAUC and NEF10 metrics, experimental percent inhibition (activity) data were binarized using a target-specific μ+2σ threshold based on the activity distribution of the PKIS1 compounds for the kinase target. Actives were defined as compounds with greater than twice the standard deviations above the mean, as noted in (1). When applying the metrics, active informer compounds were counted as true positives, whereas inactive informers did not count against the models as false positives. It should be noted that the main purpose of the informer set is to facilitate accurate activity ranking on the non-informers. However, since informer compounds represent the highest priority compounds for testing, we reward models for retrieving active informers but refrain from penalizing models for choosing inactive informers. Some baseline models that rely upon binary compound labels occasionally failed to evaluate the noninformer compounds in cases where no active informers are returned. In such cases, metric scores reflecting random ranking were assigned to the model: ROCAUC and NEF10 of 0.5 and a FASR10 score of 0.0.

FASR10 assesses a model's capacity to recognize different active chemotypes among the the top 10% of ranked compounds. The metric reflects the fraction of all active scaffolds identified on a given target within the compound set. Again, z_i is the Boolean vector of compound binary activity labels on target i for compound set J. Let O_J be the vector of chemical scaffold identifiers for compounds in J. The scaffold identifiers are arbitrary integer scaffold indices assigned to each of the generic Bemis-Murcko scaffold presented in J, as obtained using the `MurckoScaffold` module in RDKit [39, 42]. Bemis-Murcko scaffolds were made generic by stripping hydrogens, converting all bonds to single, and setting all atom types to aliphatic carbon. The unique active scaffold identifiers are the set of all non-zero values in the

Hadamard product vector:

$$C_J = \{z_i \circ O_J\} \tag{19}$$

If we then let z_i^{10} and O_J^{10} be the binary activity labels and scaffold IDs for the top 10% ranked compounds, the subset of unique active scaffolds recognized just among the top 10% of compounds is:

$$C_J^{10} = \{z_i^{10} \circ O_J^{10}\} \tag{20}$$

The fraction of active scaffolds recognized in the top 10% is:

$$\text{FASR10} = \frac{|C_J^{10}|}{|C_J|} \tag{21}$$

Note, active scaffolds were not considered *retrieved* unless an experimentally observed active member from that chemotype was in the top 10%. Cases arise where only inactive members of an active scaffold were obtained in the top 10% of the compound ranking. In such cases, the FASR10 metric does not count the chemotype as recognized.

Although the IBRs were developed for compound ranking and not necessarily as classifiers, standard classification metrics F1 and MCC were also applied for IBR performance evaluations. The scores/ranks returned by each method were converted to binary classifications using a fixed threshold across targets based on the median active fraction of the chemogenomic data: 5.5% for both PKIS1 and PKIS2. This amounts to assigning an *active* classification to the top 20 and top 23 scoring compounds in PKIS1 (366) and PKIS2 (415), respectively. As in the other metrics, inactive informers were not counted as false positives and were removed before metric calculations. Active informers, however, were counted as true positives.

Model evaluations. Performance of the models was evaluated in two stages. The first stage follows a retrospective leave-one-target-out (LOTO) evaluation scheme. Each of the 224 kinase targets in the PKIS1 target set is removed and treated as a new target of interest *i*. The PKIS1 compound activities are hidden for this target. An informer set A_i is selected for this new target, the activities are revealed for the informers, and then the model rank orders the remaining noninformers A_i^c using the informer data and in some cases data from the other 223 targets. The 9 models were evaluated in this stage using the 3 metrics described above. The second stage is a prospective evaluation of the 9 models as applied on three novel, non-human, kinase targets. In these evaluations, informer sets were generated twice for each model–once on each of the training matrices, PKIS1 and PKIS2. The remaining compounds (noninformers) from the corresponding matrix are then ranked on the two novel kinase targets using data returned for the informer sets and data within the corresponding PKIS1 or PKIS2 training matrix from which the informer set was selected. As in the retrospective PKIS1 LOTO evaluation, each model was assessed using the 3 metrics described above. However, in this prospective test on the new targets, each model was applied twice, using each of the PKIS data matrices, and therefore a total of 6 evaluations were performed on each model. We attempted to build a larger PKIS matrix by merging the PKIS1 and PKIS2 data matrices. The structure of the merged matrix, was however problematic in that the compound sets were nearly disjoint between PKIS1 and PKIS2. The resulting incomplete matrix lacks a structure that enables accurate imputation of the missing activity elements.

Code and data availability

PknB and BGLF4 screening data obtained at the UW-Carbone Cancer Center's Small Molecule Screening Facility, ROP18 data, formatted PKIS1 and PKIS2 datasets, and a Python implementation of the baseline IBR methods, evaluation metrics, and plotting procedures are available here: https://github.com/SpencerEricksen/informers. Matlab code and documentation involving the *RS* method is available here: https://github.com/leepei/informer. An R package for running *CS* and *AS* methods is available here: https://github.com/wiscstatman/esdd/tree/master/informRset.

Supporting information

S1 Fig. Nearest-neighbor sequence identity distributions for kinase domains in PKIS sets. A sequence similarity matrix (% kinase domain sequence identity) was determined for most members of the PKIS1 and PKIS2 kinase sets (mutants removed). The kinase domain sequences of targets BGLF4, PknB, and ROP18 were also included. The histograms show the distribution of nearest-neighbor sequence identities among kinase domains within the matrices (PKIS1 or PKIS2). The blue (BGLF4), red (PknB), and green (ROP18) diamonds indicate nearest neighbor sequence identities observed for these targets in the PKIS1 and PKIS2 kinase sets. BGLF4, PknB, and ROP18 do not have closely related neighbors in the sets.
(TIFF)

S2 Fig. The distribution of active compound fractions across kinase targets in PKIS1 and PKIS2. The variation in the class imbalance is even wider when a universal threshold is applied (percent inhibition) over all targets. Diamonds indicate the fraction of active compounds for BGLF4 (blue), PknB (red), and ROP18 (green) in the PKIS1 and PKIS2 compound sets.
(TIFF)

S3 Fig. Projection of PKIS1 compounds along the 3 primary components taken from PCA on their Morgan fingerprints. Target kinases EGFR and LOK are shown here as examples. PKIS1 compounds (points) are colored according to their experimental activity on the target: yellow indicates high activity (strong inhibition) and blue is low. Active compounds (exceeding threshold) have markers outlined in red. On these example kinase targets, separated regions of active chemical space can be observed.
(TIFF)

S4 Fig. PKIS1 LOTO VS performance of AS method as function of informer set size. We examined the relationship between informer set size (n_A = 9 to 28) for IBR method AS and virtual screening performance in terms of ROCAUC and NEF10 metrics.
(TIFF)

S5 Fig. PKIS1 LOTO virtual screening performance as a function of informer set size for baseline method BC_w. ROCAUC (left) and NEF10 (right).
(TIFF)

S6 Fig. PKIS1 LOTO virtual screening performance as a function of informer set size for baseline method BF_w. ROCAUC (left) and NEF10 (right).
(TIFF)

S7 Fig. Violin plots of ROCAUC stratified over four target sets (Q1-Q4) with different activity rates (Q1 lowest, Q4 highest) in PKIS1.
(TIFF)

S8 Fig. Methods effects on ROCAUC from stratified regression. Shown are 95% confidence intervals (Tukey's method) from a regression model allowing target and method to affect performance.
(TIFF)

S9 Fig. Violin plots of NEF10 stratified over four target sets (Q1-Q4) with different activity rates (Q1 lowest, Q4 highest) in PKIS1.
(TIFF)

S10 Fig. Methods effects on NEF10 from stratified regression. Shown are 95% confidence intervals (Tukey's method) from a regression model allowing target and method to affect performance.
(TIFF)

S11 Fig. Violin plots of FASR10 stratified over four target sets (Q1-Q4) with different activity rates (Q1 lowest, Q4 highest) in PKIS1.
(TIFF)

S12 Fig. Methods effects on FASR10 from stratified regression. Shown are 95% confidence intervals (Tukey's method) from a regression model allowing target and method to affect performance.
(TIFF)

S1 Table. Informer selections among PKIS1 compound set by IBR method. From 366 PKIS1 matrix compounds, 16 informer compounds were selected by each IBR method. The union of all informer compounds across 5 IBRs is listed. Informers selected by each method are indicated black dots. Informer activities are reported as normalized percent inhibition values [0, 1]. Informers considered active have activities reported in boldface.
(PDF)

S2 Table. Retrieval counts by the various methods on new kinase targets (a) PknB, (b) BGLF4, and (c) ROP18 using PKIS1 or PKIS2 matrices. The values below each of the IBR methods indicate the number of active informers observed (out of 16), the number of active compounds identified in the top 10% ranking compounds by each method, and the number of distinct active scaffolds recognized in the top 10%. The total number of experimentally determined compounds and active scaffolds is indicated in the *total* column. For a given target, molecules included in the top 10% compounds are the active informers and the top ranking non-informers comprising 10% of the set of all compounds after removing inactive informers.
(PDF)

S3 Table. Metrics evaluations against three new kinase targets (a) PKNB, (b) BGLF4, and (c) ROP18 using PKIS1 or PKIS2 matrices. IBR strategies were applied prospectively on novel kinase targets, (a) Pknb, (b) BGLF4, and (c) ROP18, which do not belong to either of the PKIS1 and PKIS2 target sets.
(PDF)

S4 Table. (a) ROCAUC, (b) NEF10, (c) FASR10, (d) F_1 score (F1), and (e) Matthew's Correlation Coefficient (MCC) in Leave-One-Target-Out Cross Validation on PKIS1. Nine IBR methods were evaluated on 224 PKIS1 targets using standard VS metrics that reflect active retrieval, ROCAUC, NEF10, F1, and MCC. FASR10 was also evaluated to reflect the chemical diversity of the actives retrieved.
(PDF)

S5 Table. *P*-values in methods comparison for PKIS1 LOTO VS. VS metrics (a) ROCAUC, (b) NEF10, (c) FASR10, (d) F_1 score (F1), and (e) Matthew's Correlation Coefficient (MCC) in Leave-One-Target-Out Cross Validation on PKIS1 are compared across 9 IBR methods evaluated on 224 PKIS1 targets. A pairwise, 2-sided Wilcoxon signed-rank test (non-parametric) was applied to calculate *p*-values. Bold font is used to indicate *p*-values that fail to pass $\alpha = 0.05$ threshold for significance when comparing IBR to baseline methods. To collectively compare all 6 baseline IBRs against each non-baseline IBR, we imposed a Šidák multiple comparison correction with 6 hypotheses. This increases the stringency of the statistical threshold applied on each of the 6 individual tests to $\alpha = 0.0085$. However, after applying this correction, non-baseline methods remained statistically superior to all baselines for all metrics except for CS when considering the ROCAUC metric and RS when compared on F1 and MCC metrics.
(PDF)

S1 Text. Pseudocode for regression selection, coding selection, and adaptive selection.
(PDF)

Acknowledgments

We thank Rob Nowak and Sebastian Raschka for feedback on the manuscript. We thank Bill Zuercher (University of North Carolina) and L. David Sibley (Washington University in St. Louis) for the ROP18 inhibition data. We also thank Yongna Xing and Vitali Stanevich for providing BGLF4 protein for assays and feedback on the manuscript.

Author Contributions

Conceptualization: Spencer S. Ericksen, Julie C. Mitchell, Anthony Gitter, Stephen J. Wright, F. Michael Hoffmann, Scott A. Wildman, Michael A. Newton.

Data curation: Spencer S. Ericksen, Gene E. Ananiev, Nathan Wlodarchak, Michael A. Newton.

Formal analysis: Huikun Zhang, Spencer S. Ericksen, Ching-pei Lee, Peng Yu, Stephen J. Wright, Michael A. Newton.

Funding acquisition: Julie C. Mitchell, Anthony Gitter, Stephen J. Wright, F. Michael Hoffmann, Michael A. Newton.

Investigation: Huikun Zhang, Spencer S. Ericksen, Ching-pei Lee, Gene E. Ananiev, Nathan Wlodarchak, Peng Yu, Scott A. Wildman, Michael A. Newton.

Methodology: Huikun Zhang, Spencer S. Ericksen, Ching-pei Lee, Stephen J. Wright, Scott A. Wildman, Michael A. Newton.

Project administration: Spencer S. Ericksen, Stephen J. Wright, F. Michael Hoffmann, Scott A. Wildman, Michael A. Newton.

Resources: Nathan Wlodarchak.

Software: Huikun Zhang, Spencer S. Ericksen, Ching-pei Lee, Anthony Gitter, Stephen J. Wright, Michael A. Newton.

Supervision: Anthony Gitter, Stephen J. Wright, F. Michael Hoffmann, Scott A. Wildman, Michael A. Newton.

Validation: Spencer S. Ericksen, Anthony Gitter.

Visualization: Spencer S. Ericksen, Michael A. Newton.

Writing – original draft: Huikun Zhang, Spencer S. Ericksen, Ching-pei Lee, F. Michael Hoffmann, Scott A. Wildman, Michael A. Newton.

Writing – review & editing: Spencer S. Ericksen, Nathan Wlodarchak, Julie C. Mitchell, Anthony Gitter, Stephen J. Wright, F. Michael Hoffmann, Scott A. Wildman, Michael A. Newton.

References

1. Lionta E, Spyrou G, Vassilatis DK, Cournia Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. Current Topics in Medicinal Chemistry. 2014; 14 (16):1923–1938. https://doi.org/10.2174/1568026614666140929124445 PMID: 25262799
2. Oprea TI, Matter H. Integrating virtual screening in lead discovery. Current Opinion in Chemical Biology. 2004; 8(4):349–358. https://doi.org/10.1016/j.cbpa.2004.06.008 PMID: 15288243
3. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational Methods in Drug Discovery. Pharmacological Reviews. 2014; 66(1):334–395. https://doi.org/10.1124/pr.112.007336 PMID: 24381236
4. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nature Reviews Drug Discovery. 2004; 3(11):935–949. https://doi.org/10.1038/nrd1549 PMID: 15520816
5. Ericksen SS, Wu H, Zhang H, Michael LA, Newton MA, Hoffmann FM, et al. Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. Journal of Chemical Information and Modeling. 2017; 57(7):1579–1590. https://doi.org/10.1021/acs.jcim.7b00153 PMID: 28654262
6. Geppert H, Vogt M, Bajorath J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. Journal of Chemical Information and Modeling. 2010; 50(2):205–216. https://doi.org/10.1021/ci900419k PMID: 20088575
7. Martin YC, Kofron JL, Traphagen LM. Do Structurally Similar Molecules Have Similar Biological Activity? Journal of Medicinal Chemistry. 2002; 45(19):4350–4358. https://doi.org/10.1021/jm020155c PMID: 12213076
8. Petrone PM, Simms B, Nigsch F, Lounkine E, Kutchukian P, Cornett A, et al. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. ACS Chemical Biology. 2012; 7 (8):1399–1409. https://doi.org/10.1021/cb3001028 PMID: 22594495
9. Cortes Cabrera A, Lucena-Agell D, Redondo-Horcajo M, Barasoain I, Díaz JF, Fasching B, et al. Aggregated Compound Biological Signatures Facilitate Phenotypic Drug Discovery and Target Elucidation. ACS Chemical Biology. 2016; 11(11):3024–3034. https://doi.org/10.1021/acschembio.6b00358 PMID: 27564241
10. Helal KY, Maciejewski M, Gregori-Puigjané E, Glick M, Wassermann AM. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. Journal of Chemical Information and Modeling. 2016; 56(2):390–398. https://doi.org/10.1021/acs.jcim.5b00498
11. Riniker S, Wang Y, Jenkins JL, Landrum GA. Using Information from Historical High-Throughput Screens to Predict Active Compounds. Journal of Chemical Information and Modeling. 2014; 54 (7):1880–1891. https://doi.org/10.1021/ci500190p PMID: 24933016
12. Bender A, Young DW, Jenkins JL, Serrano M, Mikhailov D, Clemons PA, et al. Chemogenomic data analysis: Prediction of small-molecule targets and the advent of biological fingerprints. Combinatorial Chemistry & High Throughput Screening. 2007; 10(8):719–731. https://doi.org/10.2174/138620707782507313
13. Dančík V, Carrel H, Bodycombe NE, Seiler KP, Fomina-Yadlin D, Kubicek ST, et al. Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. Journal of Biomolecular Screening. 2014; 19(5):771–781. https://doi.org/10.1177/1087057113520226 PMID: 24464433
14. Maciejewski M, Wassermann AM, Glick M, Lounkine E. Experimental Design Strategy: Weak Reinforcement Leads to Increased Hit Rates and Enhanced Chemical Diversity. Journal of Chemical Information and Modeling. 2015; 55(5):956–962. https://doi.org/10.1021/acs.jcim.5b00054 PMID: 25915687
15. Cortés-Ciriano I, Firth NC, Bender A, Watson O. Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening. Journal of Chemical Information and Modeling. 2018; https://doi.org/10.1021/acs.jcim.8b00376

16. Reker D, Schneider P, Schneider G, Brown J. Active learning for computational chemogenomics. Future Medicinal Chemistry. 2017; 9(4):381–402. https://doi.org/10.4155/fmc-2016-0197 PMID: 28263088

17. Rakers C, Najnin RA, Polash AH, Takeda S, Brown JB. Chemogenomic Active Learning's Domain of Applicability on Small, Sparse qHTS Matrices: A Study Using Cytochrome P450 and Nuclear Hormone Receptor Families. ChemMedChem. 2018; 13(6):511–521. https://doi.org/10.1002/cmdc.201700677

18. Paricharak S, IJzerman AP, Jenkins JL, Bender A, Nigsch F. Data-Driven Derivation of an "Informer Compound Set" for Improved Selection of Active Compounds in High-Throughput Screening. Journal of Chemical Information and Modeling. 2016; 56(9):1622–1630. https://doi.org/10.1021/acs.jcim.6b00244

19. Taylor R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. Journal of Chemical Information and Computer Sciences. 1995; 35(1):59–67.

20. Drewry DH, Willson TM, Zuercher WJ. Seeding Collaborations to Advance Kinase Science with the GSK Published Kinase Inhibitor Set (PKIS). Current Topics in Medicinal Chemistry. 2014; 14(3):340–342. https://doi.org/10.2174/1568026613666131127160819 PMID: 24283969

21. Drewry DH, Wells CI, Andrews DM, Angell R, Al-Ali H, Axtman AD, et al. Progress towards a public chemogenomic set for protein kinases and a call for contributions. PLOS ONE. 2017; 12(8):e0181585. https://doi.org/10.1371/journal.pone.0181585 PMID: 28767711

22. Wlodarchak N, Teachout N, Beczkiewicz J, Procknow R, Schaenzer AJ, Satyshur K, et al. In Silico Screen and Structural Analysis Identifies Bacterial Kinase Inhibitors which Act with Beta-Lactams To Inhibit Mycobacterial Growth. Molecular Pharmaceutics. 2018; 15(11):5410–5426. https://doi.org/10.1021/acs.molpharmaceut.8b00905

23. Simpson C, Jones NG, Hull-Ryde EA, Kireev D, Stashko M, Tang KL, et al. Identification of Small Molecule Inhibitors that Block the Toxoplasma gondii Rhoptry Kinase ROP18. ACS Infectious Diseases. 2016; 2(3):194–204. https://doi.org/10.1021/acsinfecdis.5b00102 PMID: 27379343

24. Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I. Predicting Drug–Target Interactions Using Probabilistic Matrix Factorization. Journal of Chemical Information and Modeling. 2013; 53(12):3399–3409. https://doi.org/10.1021/ci400219z

25. Irwin JJ, Gaskins G, Sterling T, Mysinger MM, Keiser MJ. Predicted Biological Activity of Purchasable Chemical Space. Journal of Chemical Information and Modeling. 2018; 58(1):148–164. https://doi.org/10.1021/acs.jcim.7b00316 PMID: 29193970

27. Cichonska A, Ravikumar B, Parri E, Timonen S, Pahikkala T, Airola A, et al. Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. PLOS Computational Biology. 2017; 13(8):e1005678. https://doi.org/10.1371/journal.pcbi.1005678 PMID: 28787438

26. Wassermann AM, Geppert H, Bajorath J. Ligand Prediction for Orphan Targets Using Support Vector Machines and Various Target-Ligand Kernels Is Dominated by Nearest Neighbor Effects. Journal of Chemical Information and Modeling. 2009; 49(10):2155–2167. https://doi.org/10.1021/ci9002624 PMID: 19780576

28. Janssen APA, Grimm SH, Wijdeven RHM, Lenselink EB, Neefjes J, van Boeckel CAA, et al. Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome–Inhibitor Interaction Landscapes. Journal of Chemical Information and Modeling. 2018; https://doi.org/10.1021/acs.jcim.8b00640

29. Paricharak S, IJzerman AP, Bender A, Nigsch F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data. ACS Chemical Biology. 2016; 11(5):1255–1264. https://doi.org/10.1021/acschembio.6b00029 PMID: 26878899

30. Garnett R, Gärtner T, Vogt M, Bajorath J. Introducing the 'active search' method for iterative virtual screening. Journal of Computer-Aided Molecular Design. 2015; 29(4):305–314. https://doi.org/10.1007/s10822-015-9832-9

31. Böhm HJ, Flohr A, Stahl M. Scaffold hopping. Drug Discovery Today: Technologies. 2004; 1(3):217–224. https://doi.org/10.1016/j.ddtec.2004.10.009 PMID: 24981488

32. Hu Y, Stumpfe D, Bajorath J. Recent Advances in Scaffold Hopping. Journal of Medicinal Chemistry. 2017; 60(4):1238–1246. https://doi.org/10.1021/acs.jmedchem.6b01437 PMID: 28001064

33. Elkins JM, Fedele V, Szklarz M, Abdul Azeez KR, Salah E, Mikolajczyk J, et al. Comprehensive characterization of the Published Kinase Inhibitor Set. Nature Biotechnology. 2016; 34(1):95–103. https://doi.org/10.1038/nbt.3374 PMID: 26501955

34. Arthur D, Vassilvitskii S. K-means++: The Advantages of Careful Seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. SODA'07. Society for Industrial and Applied Mathematics; 2007. p. 1027–1035. Available from: http://dl.acm.org/citation.cfm?id=1283383.1283494.

35. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

36. Kim T, Wright SJ. PMU Placement for Line Outage Identification via Multinomial Logistic Regression. IEEE Transactions on Smart Grid. 2018; 9(1):122–131. https://doi.org/10.1109/TSG.2016.2546339

37. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010; 33(1):1–22. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

39. RDKit: open-source cheminformatics software;. Available from: http://rdkit.org.

40. Rogers D, Hahn M. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 2010; 50(5):742–754. https://doi.org/10.1021/ci100050t PMID: 20426451

41. Nicholls A. What do we know and when do we know it? Journal of Computer-Aided Molecular Design. 2008; 22(3):239–255. https://doi.org/10.1007/s10822-008-9170-2

42. Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry. 1996; 39(15):2887–2893. https://doi.org/10.1021/jm9602928 PMID: 8709122