Cheminformatics tools for analyzing and designing optimized small molecule collections and libraries

Nienke Moret1,2, Nicholas A. Clark3, Marc Hafner1,2,†, Yuan Wang4, Eugen Lounkine4, Mario Medvedovic3, Jinhua Wang5,6, Nathanael Gray5,6, Jeremy Jenkins4, and Peter K. Sorger1,2

1HMS LINCS and Druggable Genome Centers, Laboratory of Systems Pharmacology, Harvard Program in Therapeutic Science, Harvard Medical School, Boston, Massachusetts 02115, USA
2Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA
3Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, Cincinnati, OH 45221
4Novartis Institutes for BioMedical Research Inc., 181 Massachusetts Avenue, Cambridge, MA 02139, USA
5Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
6Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 360 Longwood Avenue, Longwood Center 2209, Boston, MA 02115, USA

SUMMARY

Libraries of well-annotated small molecules have many uses in chemical genetics, drug discovery and therapeutic repurposing. Multiple libraries are available, but few data-driven approaches exist to compare them and design new libraries. We describe an approach to scoring and creating libraries based on binding selectivity, target coverage and induced cellular phenotypes as well as chemical structure, stage of clinical development and user preference. The approach, available via the online tool http://www.smallmoleculesuite.org, assembles sets of compounds with the lowest possible off-target overlap. Analysis of six kinase inhibitor libraries using our approach reveals dramatic differences among them and led us to design a new LSP-OptimalKinase library that

Corresponding author & lead contact: Peter K. Sorger, Warren Alpert 444, 200 Longwood Avenue, Boston, Massachusetts 02115, peter_sorger@hms.harvard.edu cc: christopher_bird@hms.harvard.edu, phone (617) 432-6902.

†Present address: Department of Bioinformatics and Computational Biology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA

CONTRIBUTIONS

Conceptualization, NM, MH, NG, JJ and PKS; Algorithm development, NM, NAC, YW and EL; Smallmoleculesuite design, NM, NAC and MH; Original Draft, NM and PKS; LINCS library assembly, NM, MH, JW, NG and PKS; Funding Acquisition, MM, NG and PKS.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DECLARATION OF INTERESTS

YW, EL and JJ and employees of the Novartis Institutes for BioMedical Research; MH is an employee of Genentech Inc. The other authors declare that they have no relevant conflicts.

Data and resource availability
All source code can be found online on www.github.com/sorgerlab/smallmoleculesuite.
outperforms existing collections in target coverage and compact size. We also describe a mechanism of action library that optimally covers 1852 targets in the liganded genome. Our tools facilitate creation, analysis and updates of both private and public compound collections.

Graphical Abstract

eTOC blurb

Despite the widespread use of small molecule libraries, few data-driven tools exist to analyze existing libraries and design new ones. Here, Moret et al. use data on binding selectivity, target coverage, induced cellular phenotype, chemical structure and phase of clinical development to analyze and design small molecule libraries.

Keywords

Screening library; mechanism of action; kinase inhibitors; cheminformatics; polypharmacology; drug repurposing; machine learning

INTRODUCTION

The size and diversity of small molecule libraries is an important design criterion: large libraries make it possible to explore a greater fraction of chemical space, but have the disadvantage that assays are high-throughput and thus, are usually highly simplified biologically. The use of small libraries makes it feasible to perform complex phenotypic assays, thoroughly study dose-response relationships, screen drug combinations and identify conditions that promote sensitivity and resistance (Fallahi-Sichani et al., 2017; Griner et al., 2014; Inglese et al., 2006; Jones et al., 2016; Lemieux et al., 2011; Niepel et al., 2013). Focused small molecule libraries (typically comprising ~ 30–3,000 compounds) are also widely used to study specific biological processes (using chemical genetic approaches) and uncover drug repurposing opportunities (Wang et al., 2016). Careful screening of focused libraries has resulted in the identification of “first-in-class” drugs (Eder et al., 2014; Moffat et al., 2014; Swinney and Anthony, 2011) such as Vorinostat (Marks, 2007) and Miglustat (Ficicioglu, 2008) and led to repurposing of (investigational) drugs such as LY500307 (Erteberel); an investigational drug in Phase Ib/II for schizophrenia (Strittmatter, 2014), after it failed to meet its endpoint in treating benign prostatic hyperplasia (Roehrborn et al.,...
2015). Despite this, few data-driven approaches exist to evaluate existing compound collections or design new ones.

Focused compound libraries often concentrate on a single, druggable, multigene family, such as kinases or ion-channels, FDA-approved drugs or the \textit{liganded genome} (which we defined as the subset of proteins in the druggable genome that are currently bound by at least three compounds having $K_i < 10 \, \mu M$); libraries against the liganded genome are commonly referred to as “Mechanism of Action” (MoA) libraries because they can be helpful in dissecting biological mechanism (Wassermann et al., 2014). Such libraries have become available through public initiatives (Keenan et al., 2017), outreach by pharmaceutical companies (Elkins et al., 2016), commercial small molecule vendors and expert-curated lists of “optimal” compounds (Arrowsmith et al., 2015; Wang and Gray, 2015a, 2015b), (e.g. \texttt{www.chemicalprobes.org}). In principle, choosing a library should be relatively simple: optimal libraries include as many bioactive and highly selective compounds as necessary to cover a target class of interest, with the possible addition of approved and investigational therapeutics. Achieving this in practice is difficult, since many compounds exhibit polypharmacology (that is, they bind to multiple protein targets) and data on polypharmacology are frequently incomplete. It can therefore be difficult to determine whether a drug-induced phenotype is due to inhibition of the nominal target, a secondary target or both (Arrowsmith et al., 2015). This also holds true in chemical genetic studies that use compounds to investigate the cellular functions of a single target of interest: in such studies, the confounding effects of off-target binding can best be mitigated by using multiple compounds with minimal off-target overlap.

In this paper, we describe a set of algorithms and software tools for optimizing the composition of focused molecule libraries and small compound collections based on binding selectivity, structural diversity, similarity in cell-based assay activities and stage of clinical development. Our approach aims to minimize the number of compounds in a library while preserving diversity and other desirable characteristics. We compare six kinase-focused libraries and find that they vary substantially in chemical diversity and the range of targets covered. We therefore assemble a new \textit{LSP-OptimalKinase} library with properties superior to any existing compound collection in terms of target coverage and compound selectivity. We also generate a compound list for an \textit{LSP-MoA} library that optimally targets the liganded genome. Since libraries are created using software, they can easily be updated as new information on small molecules and their targets becomes available.

\section*{RESULTS}

To analyze small molecule libraries we used four types of information curated in ChEMBL and other data sources: (i) chemical structure, (ii) \textit{target data}, (iii) \textit{nominal target} and (iv) \textit{phenotypic data} (from drug-treated cells). Target \textit{dose-response data} are most commonly derived from enzymatic assays that use recombinant or purified proteins at $>5$ drug concentrations, yielding curves from which $K_i$ or IC$_{50}$ values can be derived. Target \textit{profiling data} primarily involves assays with large panels of recombinant proteins, or proteins present in cell-free extracts, and yields information on binding to many different targets, which is often expressed as percent activity against each target (see for example \textit{Cell Chem Biol.} Author manuscript; available in PMC 2020 May 16.)
DiscoverX KINOMEscan (https://www.discoverx.com/home) and Kinativ (http://www.kinativ.com/) (Anastassiadis et al., 2011; Davis et al., 2011; Karaman et al., 2008) “Nominal target” is the target most commonly associated with a drug (e.g. the “BRC-ABL inhibitor imatinib”) but it may not be the highest affinity nor the only biologically relevant target. “Phenotypic data” comprise experimental data from cell-based assays in which drug activity is measured using a morphological, biochemical, genetic or functional readout but drug-target engagement is not determined directly (e.g. drug-induced changes in proliferation, apoptosis, migration, immune-cell activation, transcription factor translocation, etc.).

The data sources used in this work comprised the ChEMBLV22_1 database (Bento et al., 2014), which draws on data published in the scientific literature, patents, PubChem bioassays, FDA approval packages and other resources (Gaulton et al., 2017); kinome-wide screens from the International Centre for Kinase Profiling (http://www.kinase-screen.mrc.ac.uk); kinome-wide screens from LINCS (Keenan et al., 2017); and in-house curation of nominal targets (Figure S1). To correctly combine data for a single compound under different names (e.g. OSI-774, Erlotinib and Tarceva®) we matched chemical structures using the Tanimoto similarity of Morgan2 fingerprints (hereafter “structural similarity” or “Tc”) (Rogers and Hahn, 2010).

Our analysis of focused libraries involves six widely available kinase inhibitor libraries with public annotations (Table S1): (i) the SelleckChem kinase library (abbreviated SK), a set of 429 kinase inhibitors assembled by the commercial supplier SelleckChem; (ii) the “Published Kinase Inhibitor Set” (PKIS), a 362 compound collection assembled by Glaxo-Smith-Kline that pioneered the concept of industry-sponsored ‘open-source’ drug discovery; (iii) the Dundee compound collection (Dundee), a set of 209 kinase inhibitors screened for biochemical activity at the University of Dundee; (iv) the EMD kinase inhibitor collection (EMD), a set of 266 kinase inhibitors sold by Tocris Bioscience; (v) the HMS-LINCS small molecule collection (LINCS), a set of 495 small molecule probes and drugs focused on kinase inhibitors that was assembled by the NIH-funded HMS-LINCS center (which is led by the senior author on this paper); and (vi) SelleckChem Pfizer licensed collection (SP), a set of 94 molecules licensed by Pfizer for sale by SelleckChem (see Table S1 for references and Web links). In aggregate these libraries contain 1417 unique compounds all of which could be mapped onto ChEMBL identifiers (Methods; Tables S2 and S3C).

**Chemical similarity within libraries**

With respect to composition, we found that the LINCS and SK libraries shared ~50% of their compounds, making them the most similar among the six libraries examined (Figure 1A). In contrast, 350 of 362 compounds in the PKIS library were unique. To visualize structural diversity, we plotted chemical similarity against itself; the presence of highly similar compounds (analogues) gives rise to off-diagonal clusters. The LINCS library generated few such clusters whereas the PKIS library was dominated by them (the PKIS library was designed to have such clusters of structurally similar compounds; Figure 1B). By way of illustration, a three-compound cluster of LINCS analogues comprised the CDK inhibitor Seliciclib, and two close relatives, Olomoucine II and (S)-CR8 (Figure 1B bottom).
Seliciclib progressed to phase II for nasopharyngeal carcinoma, but failed to meet its primary endpoint of improving progression free survival (Khalil et al., 2015; Yeo et al., 2009), while the other compounds in the same cluster are reported to have improved cellular potency and selectivity (Kryštof et al., 2002; Oumata et al., 2008). A nine-compound PKIS cluster contained a set of close analogs, none of which have progressed into clinical development (Figure 1B, bottom). By scoring the frequency and sizes of clusters having structural similarity ≥0.7 we find that the LINCS and Dundee collections are the most structurally diverse, PKIS the least, and SK, SP and EMD in-between (Figure 1C). This may arise because LINCS, Dundee and SK compounds were drawn from different drug discovery campaigns across multiple companies and research groups, whereas the PKIS library was derived from structure-activity studies inside a single company. The LINCS and SelleckChem kinase libraries also benefit from having been more recently assembled.

**Phenotypic similarity and clinical development**

To assess the diversity of phenotypes elicited by different compounds, we constructed phenotypic fingerprint (PFP) vectors that quantify activity across a wide range of phenotypic assays curated in ChEMBL. The length of PFP vectors ranged from 5–58 entries (median = 11) for pre-clinical molecules and investigational drugs and 20–870 assays (median =440) for approved drugs. For example, the vectors for imatinib and nilotinib had 303 assays in common and yielded a Pearson correlation of r=0.82, whereas imatinib and dasatinib had 426 assays in common and yielded a Pearson correlation of r=0.16 (Figure 2A). These correlation values mirror the distinct binding conformations that the drugs assume when binding the BCR-ABL oncogene: imatinib and nilotinib are type II inhibitors, locking the ABL kinase domain in its inactive conformation (Schindler et al., 2000) while dasatinib is a type I inhibitor and targets ABL in its active conformation (Kimura et al., 2005; Shah et al., 2004).

We find that the LINCS and SK libraries have the lowest correlations when all pairs of PFP vectors are considered, whereas the PKIS and SP libraries have the highest correlation and are therefore the least diverse (Figure 2B). A potential complication in this comparison is that the number of phenotypic assays reported in ChEMBL varies from compound to compound and also, on average, from library to library. Thus, libraries might appear to have lower phenotypic diversity simply because they are insufficiently annotated. This issue is particularly acute for compounds in the PKIS library, of which few are extensively annotated.

Gene expression profiles from cells exposed to small molecules provide an independent means of measuring a phenotype. A large number of such profiles (~ 1.5 × 10^6) is available through the Connectivity Map (CMAP; http://clue.io/cmap) (Duan et al., 2016; Subramanian et al., 2017). For compounds in the LINCS library we compared correlation coefficients for pairs of PFP vectors with correlation in CMAP L1000 profiles (using level 5 characteristic direction vectors (Niepel et al., 2017)). The majority of compounds (53%) exhibiting noncorrelated PFPs showed noncorrelated CMAP profiles (|r| < 0.2 CMAP, |r| < 0.25 PFP). However, it is noteworthy that CMAP profiles and PFP profiles are relatively poorly correlated overall: only 3%–7% of pairs with positive correlations in PFP (depending on
threshold used to define correlation, see methods) showed correlated CMAP profiles and none of the pairs with anticorrelated PFP had anticorrelated CMAP profiles. For some compound pairs, the discrepancy between CMAP and PFP was not surprising. For example, when pairs of compounds targeting PI3K-MTOR-AKT signaling were compared, CMAP profiles were more likely to be correlated than PFP values, which is consistent with data showing that engaging similar targets (a strong contributor to CMAP profile) in different tumor cell types elicits cell-type specific phenotypes (captured by PFP value) (Niepel et al., 2017). However, further research is required to understand why PFP and CMAP are not more similar overall.

The stage of clinical development is another criterion on which compounds in a library can be evaluated, since the study of approved or investigational therapeutics has intrinsic value. It should be noted, however, that clinical grade compounds do not necessarily represent the most selective or structurally diverse compounds. LINCS has the greatest number of approved drugs (n=67; 14% of total); SK has the greatest number of compounds that are currently or have previously been tested in phased clinical trials (n=137; 32%); SP has the highest proportion of approved drugs (n=54; 57%) and PKIS the lowest (0%) (Figure 2D). Below we describe how clinical stage, selectivity and compound diversity can be balanced against each other during library assembly.

**Compound selectivity and library coverage**

Compounds are commonly described with respect to their nominal target (e.g. a “BRAF-inhibitor”, in the case of dabrafenib). However, this does not mean that the nominal target is the highest affinity target or the most biologically or clinically relevant. Crizotinib, for example, was developed as an inhibitor of the cMET receptor, but was subsequently found to target ROS1 and ALK kinases: cell culture studies showed that apoptosis is induced specifically in NSCLC lines carrying an EML4-ALK translocation (Poon and Kelly, 2017) and the FDA initially approved the drug for tumors with this type of genetic aberration (Kazandjian et al., 2014). In general, it is difficult to determine which target is most relevant for a specific phenotype, nor is it always clear that any single target – rather than a collection of targets – is responsible for biological activity. We therefore used the selectivity score developed by Wang et al (Wang et al., 2016), which assesses selectivity for each compound-target interaction, rather than for a compound overall, while also accounting for data bias.

The selectivity score evaluates compound-target pairs based on the log-distribution of on-target affinities relative to the log-distribution of off-target affinities; this yields a selectivity score for every target against which a compound has been tested. Selectivity score measures the magnitude of difference in the first quartile of the on-target and off-target distributions (q1) and determines the rank sum p-value to assess significance. The selectivity score accounts for the fact that many compounds have only been evaluated on a relatively small number of targets, which can falsely give the impression that a compound is highly selective. For example, in our study, selectivity is judged to be higher for compound A than compound B if A has been tested against a large number of targets and found to not bind most of them even if B has a higher affinity for a particular target of interest but has not be tested against
alternative targets. It should be noted that like other features used to assess compounds in this study, selectivity score changes over time as additional data become available.

*Selectivity score* is a continuous variable, however, for convenience we also assigned descriptive labels to different classes of compound-target interaction. Most Selective (MS) interactions meet four criteria: (i) the difference in q₁ values computed for the distributions of on- and off-target data is not less than 100-fold, (ii) the distributions of on-target and off-target binding affinities differ with p-value below 0.1 (iii) the compound has at least four-fold more off-target than on-target affinity measurement (so that data bias is below 20%), (iv) at least two published reports establish that the affinity for drug-target interaction is less than 100 nM. Semi-Selective (SS) interactions have slightly less stringent criteria: (i) the difference in q₁ values is not less than 10-fold, (ii) the p-value for on and off-target binding affinities is below 0.1 (iii) at least four publication report affinity under 1 µM (because the affinity cut-off is less stringent than for interactions meeting MS criteria, more data-points are needed to establish confidence in binding). Poly-Selective (PS) interactions are those for which: (i) q₁ values for on- and off-target binding are similar, (ii) on-target q₁ is under 9 µM (iii) data bias is under 20%. Unknown (UN) interactions are those for which the compound has not been sufficiently probed for off-target binding but fulfills the PS requirements with respect potency and affinity for the nominal target.

The requirement for a multi-parametric approach is exemplified by the tool compound PP121, whose nominal targets are BCR/ABL1 and PDGFR, and dasatinib, an approved drug whose nominal targets are BCR/ABL1, Src family kinases and PDGFR. dasatinib has more known targets with an affinity between 1nM and 1 µM than PP121 (Figure 3A; red circles), but it has also been tested on a substantially larger number of kinases. In many cases, it exhibits no inhibitory activity against these off-targets (Kᵢ > 10 µM, black). Thus, the selectivity score for dasatinib’s ABL1 binding (score = 1.1) is substantially higher than that of PP121 (0.44). We found that the LINCS and SK libraries had the largest number of kinases with MS interactions to any compound in the library and the PKIS library the fewest (Figure 3B).

**Evaluating compound similarity on binding spectrum**

To compare compounds based on all known targets and known non-targets, we sought a means to summarize all available target data on a compound, including dose-response data, profiling data and curated annotations from the literature (e.g. “drug X was found to not inhibit …” or “was equipotent for …”). The resulting Target Affinity Spectrum (TAS) vectors summarize binding information from multiple assay formats weighted for the degree of evidence for high affinity binding and also for non-binding (Kᵣ > 10 µM). TAS vectors contain several best-guess weighting parameters and are therefore less precise than selectivity scores, but they provide a more complete picture of available data (Figure 4A, Figure S2, S3). For the LINCS library, TAS vectors varied in length from a median of 10 features for probe compounds (range 1–394) and 15 for approved drugs (range 1–420), (Wilcoxon rank-sum test p=0.07). In no cases did a TAS include assertions for all 545 human kinases (Eid et al., 2017), let alone for all ~3000 members of the druggable genome (Hopkins and Groom, 2002), which illustrates the sparseness of target annotation. TAS
similarities for all compound pairs were calculated using a weighted Jaccard similarity metric (see methods). We found that TAS similarity was robust to at least 20% random variation in user-assigned parameters (Figure S4).

We compared TAS vectors computed from ChEMBLv22_1 and other resources to data from a systematic assessment of kinase-inhibitor binding recently published by Klaeger et al (Klaeger et al., 2017) that uses whole-cell lysates and mass spectrometry to profile compound-target binding (ChEMBL v22_1 does not yet include this data). Surprisingly, the target coverage of the two datasets only partially overlapped, with a typical LINCS compound having one-third of its targets annotated in both datasets and one third in only one of the two datasets (Figure S5A). The reasons for differences in target spectrum as curated by ChEMBL and measured by Klaeger et al. are not yet clear and may relate to the range of targets evaluated in each case, but for targets annotated in both sources, 90% of the target-affinity assertions were congruent (Figure S5B).

We also asked whether knowledge of a molecules’ nominal target was informative about target spectrum by comparing TAS similarities for compounds with the same or different nominal targets. Remarkably, we found that the distribution of TAS similarities was very similar for compounds having the same nominal targets and for compounds having different nominal targets (Figure 4B). Thus, relatively little information about target spectrum can be discerned from nominal target alone. For example, consider two pairs of compounds: (i) PD-173074 and cabozatinib and (ii) BMS-777607 and ponatinib (Figure 4C). PD-173074 and cabozatinib are structurally dissimilar (structural similarity 0.13), but overlap in nominal targets, which are KDR and FGFR1 receptors for PD-173074 and KDR for cabozatinib (black arrows in Figure 4C). These two compounds have non-overlapping secondary targets and thus, their TAS similarity is 0.19. In the case of BMS-777607 and ponatinib (structural similarity 0.12), the nominal targets are different (MET and ABL1, respectively), but the secondary targets are very similar (Figure 4C right panel) and their TAS similarity is therefore 0.82. We propose that TAS, or a functional equivalent, rather than nominal target be more widely used when discussing and analyzing the effects of compounds on cellular physiology.

**Relationship among structural, phenotypic and target similarity**

For many compounds, data on target spectrum and induced cellular phenotype are so sparse that PFP or TAS vectors are not computable. We asked whether, in these cases, structural similarity could be used as a surrogate to assess diversity. We sought to a cutoff in structural similarities for which the great majority of compound pairs have TAS and PFP vectors that are noncorrelated (and sufficient data is available for TAS and PFP vectors to be computed). In general, we found setting a structural similarity threshold of $T_c \leq 0.2$ was an effective surrogate for diversity when PFP or TAS vectors could not be computed (Figure 5A). However, the converse was not true: among compound pairs with high structural similarity ($T_c > 0.7$) few had high PFP correlation ($<10\%$ for PFP correlation $\geq 0.9$ and $<30\%$ for PFP correlation $\geq 0.6$; Figures 5B and S5B). These findings were also true of a larger set of 9476 kinase inhibitors and 9512 GPCR inhibitors from ChEMBL (Figure 5C and 5D). For compounds in the LINCS library (which has high structural diversity), we found that pairs
with dissimilar TAS vectors are low in structural similarity (Figure 5E). This was also true for a subset of compound pairs in the PKIS library that shared the same core structure: compounds with low target similarity are low in structural similarity (≤0.2 Figure 5F). Once again, the converse did not appear to be true: structural similarity was not predictive of target similarity. Thus, structural, phenotypic and target similarity contain non-redundant information about pairs of small molecules. Moreover, when data on target binding or phenotypic effects is missing, it is possible to exploit our finding that structures with very low similarity (≤0.2) have the highest likelihood of being diverse at the phenotypic and target levels; a finding that is congruent with previous guidelines for the use of chemical probes (Arrowsmith et al., 2015).

Designing an improved kinase inhibitor and a mechanism of action library

To enable construction of new libraries and compound collections using the methods described above, we created the LibraryR software tool and deployed it as a web application using R-Shiny (http://www.smallmoleculesuite.org). For a user-specified set of target proteins (identified as HUGO gene symbols), LibraryR mines ChEMBL to create libraries containing (i) two compounds maximally specific for each gene of interest but having structural similarity ≤0.2, so that overlap in secondary targets is minimized (ii) all available compounds at a user-specified clinical stage (e.g. clinical phases 1–3 or approved) that also bind to the genes of interest with user-specified affinity (default Ki ≤1µM). LibraryR also calculates the selectivity score for each compound and assigns each compound to a selectivity class (MS, SS, PS, UN; Figure 6A). We used LibraryR to construct an improved version of the LINCS kinase library (the Laboratory of Systems Pharmacology LSP-OptimalKinase library; Table S5) and a mechanism of action library that optimally targets the liganded genome (LSP-MoA; Table S5).

Both the LSP-OptimalKinase and LSP-MoA libraries are tiered based on selectivity and clinical phase to allow sub-libraries of different sizes to be assembled (Figure 6B). Tier A is the minimal library and contains only those compounds (i) binding the specified list of genes with MS selectivity and (ii) FDA approved drugs binding more strongly than the affinity cutoff (which did not result in library redundancy, see Figure S6). Tier B adds compounds with SS selectivity that target genes not covered by MS selectivity plus all compounds that bind the genes of interest and are in clinical development (clinical phase I-III). Tier C adds compounds from PS and UN specificity classes to maximally cover the user-specified list. Tier D includes all compounds historically included in a pre-existing library or present for other reasons (e.g. as positive and negative controls; Figure 6B). Inclusion of Tier D molecules facilitates historical comparisons, collection of phenotypic fingerprints and adds tool compounds; it also increases the chance for serendipitous discovery, which scales roughly with library size. Depending on the number of tiers selected, the LSP-OptimalKinase library ranges in size from 130 compounds to 590 compounds.

The use of LibraryR substantially increased target coverage and decreased the number of compounds in LSP-OptimalKinase relative to the LINCS Kinase Library we had assembled by hand several years ago. The 130 compounds in Tier A of LSP-OptimalKinase covers 63 kinases with two compounds that have a structural similarity ≤0.2 and targets a total of 136
kinases with $MS$ selectivity as compared to 117 kinases by the 495 compounds in the original LINCS library. Adding 127 Tier B compounds extends $MS$ and $SS$ coverage to 37 additional kinases (Figure 6E, Figure 3C) while including many more compounds in clinical development (Figure 6F). However, the further addition of 343 Tier C and D compounds extends coverage to only 10 additional kinases at $PS$ and $UN$ selectivity. Thus, computation-driven library design can reduce the size (and cost) of screening libraries and the studies that use them, while increasing target coverage and maintaining library diversity (Figure S6).

The LSP-MoA library is designed to target the 1852 members of the liganded genome (Table S4) about 12% of which can be bound with $MS$ selectivity, 10% maximally with $SS$ selectivity, 44% maximally with $PS$ or $UN$ selectivity; 34% of targets are bound only by compounds that are known to have a higher affinity for another target (Figure 6C). In the case of targets bound with $UN$ selectivity, 61 involved relatively tight binders (compounds with $K_d < 100 \text{ nM}$) for which little selectivity data is available. Such compound-target pairs would benefit disproportionately from more data on possible off-targets (e.g. obtained by profiling). We analyzed LSP-MoA for its coverage of five therapeutic target classes (Harding et al., 2018): kinases, G-protein-coupled receptors (GPCRs), ion channels (ICs), nuclear hormone receptors (NHRs), and transporters (Table S4). Kinases were by far the best covered with 136 of 545 kinases (25%) at $MS$ selectivity, followed by 28 of 416 GPCRs (7%), 2 of 49 NHRs (4%), 2 of 285 ICs (0.7%) and 1 of 510 transporters (0.1%; Figure 6D). The same trend was also observed for the understudied (‘dark’) ligandable genome as recently curated by the NIH (https://commonfund.nih.gov/idg Table S4).

**DISCUSSION**

The ideal focused library of annotated, biologically active small molecules comprises the minimal number of compounds needed to cover a gene set of interest with high selectively while optionally including additional approved and investigational drugs. Achieving optimal coverage is difficult in practice since many small molecules bind multiple targets and data on target spectra are nearly always incomplete and subject to poorly understood experimental confounders. Moreover, in many cell-based screening projects, the focus is not on targets per se but instead on induced cellular phenotypes. In this paper, we develop a set of criteria for evaluating and constructing small molecule libraries based on binding selectivity, target coverage and cell-based phenotypes as well as chemical structure and stage of clinical development.

As a rule, different compounds have been tested against different sets of targets, introducing uncertainty about their target spectrum: one compound can appear more selective than another simply because it has been tested against fewer possible targets. We counterbalanced this uncertainty by applying a selectivity score that takes data bias into account and incorporated both evidence of binding and known non-binding. Whenever possible, library tier A (the tier that best balances selectivity and coverage) includes two molecules which differ in structure, thereby minimizing the likelihood of known or unknown off-target effects. This is analogues to using multiple siRNAs or guide RNAs in RNAi or CRISPR screens. Overall, we find that data on structure, target spectrum and induced phenotypes are
non-redundant, demonstrating that compounds should only be labeled as being functionally ‘similar’ or ‘different’ when evaluated on multiple criteria.

We have codified our approach to scoring and assembling compound collections in three web-based applications: SelectivitySelectR, SimilaritySelectR and LibraryR, all of which draw data from ChEMBL and other user-specified sources. For a given list of targets, libraries are constructed so that each target is ideally bound by two high-affinity compounds with diverse target spectra and induced cellular phenotypes. In cases in which phenotypic data are too sparse to be informative, we have found that setting a bar for structural diversity ($T_c \leq 0.2$) is an effective way of ensuring functional diversity.

This is in agreement with the general consensus (Arrowsmith et al., 2015; Workman and Collins, 2010) although the cutoff we set for chemical similarity is relatively stringent. Approved therapeutics, beyond two per target, are included in subsequent tiers of the library because such compounds have intrinsic value for mechanism of action, translational and drug repurposing studies. The bottom tiers in a library attempt to cover additional targets at lower levels of selectivity, including tool compounds in active development and historical compounds needed for comparison to previous libraries.

When six widely-available existing kinase inhibitor libraries were evaluated, we found that the LINCS and SK libraries had the greatest diversity, broadest coverage of the kinome and the largest number of approved and investigational drugs. These were also the largest libraries analyzed, which increases the cost of creating and using them, particularly when complex dose-response or phenotypic assays are involved. The Dundee, EMD and SP libraries were substantially smaller (19% – 54% the size of LINCS), but underperformed in one or more metrics. The PKIS library scored poorly with respect to structural and phenotypic diversity and number of approved drugs, but did have the largest number of compounds not included in any other collection; it must also be recognized as a pioneer in industry-sponsored open drug discovery. When we used LibraryR to design a new LSP-OptimalKinase library we found that a collection of 256-compounds (Tiers A plus B; see Figure 6) outperformed all existing kinase libraries with respect to selectivity, target coverage, structural diversity and number of approved and investigational drugs. This library is also smaller than all but the SP library, which has substantially poorer coverage of the kinome. While the LSP-OptimalKinase collection represents an optimized way to construct a compound collection using commercially-available compounds, it is interesting that only 12% of the kinome (63 proteins) can be targeted with two structurally diverse compounds that have MS selectivity (as compared to 8% by the LINCS and <1% by the PKIS libraries); we therefore estimate that a truly optimal kinase library would consist of approximately 1,000 compounds. Kinases are one of the most heavily studied classes of protein targets, particularly in oncology, and it is noteworthy that a substantial portion of the kinome nonetheless remains unaddressed with available chemical tools (one motivation for the NIH Illuminating the Druggable Genome program (Oprea et al., 2018)).

Because our approach is data-driven and considers only public information, some differences exist between our top picks and first choice compounds suggested by others (Arrowsmith et al., 2015; Wang and Gray, 2015a, 2015b). For example,
www.chemicalprobes.org lists BI-D1870 as a selective inhibitor of the RPS6KA1 kinase, while we score this drug-target pair as being of unknown selectivity. In the published literature D1870 has only been tested against three targets, which LibraryR considers insufficient evidence of selectivity. However, LibraryR has an option to add compounds from expert-opinion lists and then rank them relative to all other compounds in a collection.

We also generated a list for a 2026-compound “mechanism of action library (LSP-MoA) that optimally covers the liganded genome and should be of general utility in semi-focused screening campaigns. We found that kinases were the target class best covered by this library, GPCRs were a distant second and many other multi-protein families had few if any selective inhibitors. This further emphasizes the relative sparsity of selective coverage of targets considered ‘ligandable,’ at least by compounds whose structures are in the public domain. However, our approach to analyzing and designing compound libraries is data-driven so that the MOA library can be steadily improved as new data become available.

We expect the methods described here to assist in a variety of informatic, chemical biology and drug repurposing studies. For cell-based experiments, our tools can optimize the search for multiple, selective, and structurally diverse compounds against one or more target(s) of interest, thereby reducing the likelihood that off-target effects go unrecognized (in many cases, this is a 2–8 compound collection in many cases). Machine learning projects should also be able to make use of target affinity spectrum to investigate molecular profiling data on drug-treated cells, particularly when multiple compounds with the same or similar nominal targets are compared. Extension of the current work will also be important; for example, it should be possible to further customize compound collections using large-scale profiling data of the cell-types on which the screen will be performed (e.g. the Connectivity Map (Lamb et al., 2006; Subramanian et al., 2017) and Cancer Cell Line Encyclopedia (Barretina et al., 2012)). Finally, we hope that the ready availability of data on the full spectrum of compound activities will replace nominal target alone as a means to describe mechanism of action.

STAR METHODS

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by Peter Sorger, peter_sorger@hms.harvard.edu (please copy Chris Bird: chris_bird@hms.harvard.edu).

Method Details

Compute chemical identity and similarity—Compounds from the different libraries were matched to each other and their ChEMBL equivalent based Tanimoto similarity of their Morgan fingerprints (at 2Å). Those compounds that had multiple matches with 100% similarity were curated manually. See script “01_chemical_identity&similarity.py”

Collecting data from ChEMBL—A local copy of ChEMBL V22_1 was installed. Both phenotypic and target data were collected from the ACTIVITIES table. The two datatypes were distinguished based on assay annotation detailed in script
“02_collecting_data_chembl.r”. Target identifiers were converted to ENTREZ gene identifiers using ‘gene_symbols’ from the COMPONENTSYNONYMS table and targets with multiple or no matches were manually checked. See Table S2 for mapping. Maximum clinical development stage was obtained from the MOLECULE_DICTIONARY table.

**Calculating PFP correlation**—Results from qualifying phenotypic screens were normalized to robust z-scores (r-scores). For each compound pair, Pearson’s correlation was calculated on the subset of assays in which both compounds had been measured and at least one of the compounds was active (i.e. r-score ≥3 or r-score ≤−3); correlation for pairs that had less than five of such assays were excluded. See script “03_calulating_PFPsimilarity.r”.

**Calculating CMAP profile correlation**—Level 5 of the CMAP data was downloaded from GEO accession number GSE92742 on May 1st, 2017 and filtered to retain only compounds in the LINCS library. For each pair of compounds, all Pearson correlations were calculated between perturbations measured in the same cell line, at the same time point, and with same treatment concentrations (see script ‘04_import_gctx_allconc.py’). The final value reported in Figure 2C is the median distance for each pair of drugs.

**Comparing CMAP correlation to PFP correlation**—CMAP correlations and PFP correlations were merged. Noncorrelated was defined |r| < 0.2 for CMAP and |r|<0.25 for PFP. Several cutoffs for correlated were explored; CMAP r ≥0.2 or 0.3; PFP r ≥0.25 or 0.5 or 0.7.

**Calculating selectivity score**—The selectivity score for compound-target interactions was calculated according to an adapted version of Wang et al (Wang et al., 2016), so that:

\[
\text{Selectivity score} = \text{Differential IC}50 + \text{Pvalue} + \text{Assay bias}
\]

with:

\[
\text{Differential IC}50 = \log_{10}\left(\frac{\text{IC}50\text{ of target}}{\text{IC}50\text{ on target}}\right)
\]

\[
\text{Pvalue} = -\log_{10}(\text{rank}_\text{sum}_\text{Pvalue}) \quad \text{where rank}_\text{sum}_\text{Pvalue} \text{ is capped at } 10^{-15}
\]

\[
\text{Data bias} = \frac{1}{N \text{ measurements on Target}}
\]

Selectivity classes, based on differential IC50, Pvalue and Data bias, were determined as outlined in Figure 3B. See script “05_calulating_selectivity.r”.

*Cell Chem Biol.* Author manuscript; available in PMC 2020 May 16.
Calculating TAS similarity—The assignment of weights in the TAS vector was customized to each assay-type. In general, a weight of 1 denotes the highest confidence in binding and a weight of 10 denotes the highest confidence in confirmed non-binding. Dose response data was obtained from ChEMBL, DiscoverX and Kinative kinome profiles. Binding values were given a weight 1 if \(K_d < 100\text{nM}\), a weight 2 if \(100\text{nM} \leq K_d < 1\mu\text{M}\), a weight 3 if \(1\mu\text{M} \leq K_d < 10\mu\text{M}\), and weight 10 if \(K_d \geq 10 \mu\text{M}\). Annotations of nominal target were obtained from manual literature searches) and was assigned a weight of 2. One-dose data from profiling methods was obtained from weighted based on the retrospective probability of binding as explained in text.

Five different data sources were used to compose the TAS: (i) dose-response data from Chembl v22_1 (see above), (ii) dose-response data from DiscoverX kinomescan, (iii) profiling data from DiscoverX at doses 0.1uM; 1uM; 10 uM; (iv) profiling data from Dundee (http://www.kinase-screen.mrc.ac.uk/kinase-inhibitors) at doses 0.01uM; 0.1uM; 1uM; 10uM, (v) curated nominal target annotations. Targets were normalized to gene_ids using gene symbol matches and manual conversions. Binding enrichments for datasources iii and iv were calculated per concentration per source see Figure S2, Figure S3, and Table S3 for conversion results. Nominal targets were extracted from manual searches in pubmed and by mining vendor information. Binding assertions were assigned as illustrated in figure 4A. TAS vectors of compound pairs were joined and entries where one of the compounds had an assertion \(\leq 3\) were kept. If the joined vector had a length \(\geq 5\), a weighted Jaccard similarity coefficient was calculated with the formula:

\[
\text{Similarity}(C_1, C_2) = \frac{\sum_{k} \min(C_1, C_2)_{k}}{\sum_{k} \max(C_1, C_2)_{k}}
\]

see Table S3 for full TAS vectors.

Assembling new small molecule libraries—LSP-OptimalKinase was developed to target human kinases (Table S4), LSP-MoA was developed to target the liganded genome (Table S4). In the selectivity arm, the compound with the highest selectivity score was selected for each target. Then, a second compound within that same selectivity class (MS, SS, PS, UN) and a structural similarity \(\leq 0.2\) was selected. For targets with no second compound in MS class with sufficient structural dissimilarity, the SS class compounds were sourced to supply the second compound. If a target was covered by MS compound pair, SS, PS and UN compounds for the same target were omitted (unless included for clinical development); If target was covered by SS compounds, PS and UN compounds for same target were omitted; If target was covered by PS compounds, UN compounds for the same target were omitted. The clinical development arm was assembled by selecting all molecules with max_phase 1–4 in MOLECULE_DICTIONARY table that had an affinity \(<1000 \text{nM}\) for one of the targets aimed at. Tiers were developed as illustrated in figure 6B. The members of molecular target classes for which the LSP-MoA library was analyzed were obtained from http://www.guidetopharmacology.org/download.jsp (accessed Dec 15th, 2017).
The Small Molecule Suite user interface—The small molecule suite comprises the applications SelectivitySelectR, SimilaritySelectR and LibraryR. For SelectivitySelectR, the selectivity score and mean affinity were calculated for all small molecules in the LINCS compound collection for each target their affinity was measured. In the application, users can select a target of interest and see the mean affinity and selectivity for all LINCS compounds with a known affinity for this target. All results can be downloaded for further use. For SimilaritySelectR, the structural similarity, TAS similarity and PFP similarity was calculated for all compound pairs in LINCS. In the application, users can select a reference molecule of interest to identify molecules that are similar or dissimilar in structure, target spectrum or induced phenotype. The application allows users to download information on the reference compound plus three additional compounds. For LibraryR, compounds were selected based on selectivity, clinical development stage and expert opinion status. To enable selection for compounds on selectivity the selectivity score was calculated for all compounds-target pairs in ChEMBL V22_1 for which the target was a member of the liganded genome and the compound was known to bind at least one target in the liganded genome with $K_i \leq 10\mu M$. For each target, the compound with the highest selectivity score and another compound with a structural similarity $\leq 0.2$ within the same selectivity class (MS, SS, PS or UN). To select compounds on clinical development stage, all compounds with $K_i \leq 10\mu M$ against any target in the liganded genome in development phase I-IV were selected. Users can set stricter requirement for affinity when composing the library (default $K_I \leq 1\mu M$). Compounds curated by expert opinion were obtained from www.chemicalprobes.org (Arrowsmith et al., 2015) (only those rated with 4/4 stars) and a recent publication by Wang & Gray (Wang and Gray, 2015a, 2015b).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by NIH LINCS and IDG grants U54-HL127365 (NM, PKS and NG), U24-DK116204 (NM, PKS) and U54-HL127624 (NAC and MM) and by the Novartis Institutes for BioMedical Research (YW, EL and JJ). We thank Dr. Verena Becker for creating the literature annotation table for compounds in LINCS, the ICCB-Longwood Screening Facility at Harvard Medical School for compound management and Dr. Christophe Benoist for mentorship and guidance.

REFERENCES

Anastassiadis T, Deacon SW, Devarajan K, Ma H, and Peterson JR (2011). Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat. Biotechnol 29, 1039. [PubMed: 22037377]

Arrowsmith CH, Audia JE, Austin C, Baell J, Bennett J, Blagg J, Bouatra C, Brennan PE, Brown PJ, Bunnage ME, et al. (2015). The promise and peril of chemical probes. Nat. Chem. Biol 11, 536–541. [PubMed: 26196764]

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607. [PubMed: 22460905]

Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, et al. (2014). The ChEMBL bioactivity database: an update. Nucleic Acids Res 42, D1083–90. [PubMed: 24214965]
Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, and Zarrinkar PP (2011). Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol 29, 1046–1051. [PubMed: 22037378]

Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, Readhead B, Tritsch SR, Hodos R, Hafner M, et al. (2016). L1000CDS2: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl 2.

Eder J, Sedrani R, and Wiesmann C (2014). The discovery of first-in-class drugs: origins and evolution. Nat. Rev. Drug Discov 13, nr4336.

Eid S, Turk S, Volkamer A, Rippmann F, and Fulle S (2017). KinMap: a web-based tool for interactive navigation through human kinome data. BMC Bioinformatics 18, 16. [PubMed: 28056780]

Elkins JM, Fedele V, Szklarz M, Azezee KRA, Salah E, Mikolajczyk J, Romanov S, Sepetov N, Huang X-P, Roth BL, et al. (2016). Comprehensive characterization of the Published Kinase Inhibitor Set. Nat. Biotechnol 34, 95. [PubMed: 26501955]

Fallahi-Sichani M, Becker V, Izar B, Baker GJ, Lin J-R, Boswell SA, Shah P, Rotem A, Garraway LA, and Sorger PK (2017). Adaptive resistance of melanoma cells to RAF inhibition via reversible induction of a slowly dividing de-differentiated state. Mol. Syst. Biol 13, 905. [PubMed: 28069687]

Ficicioglu C (2008). Review of miglustat for clinical management in Gaucher disease type 1. Ther. Clin. Risk Manag 4, 425–431. [PubMed: 18728838]

Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutovo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res 45, D945–D954. [PubMed: 27899562]

Griner LAM, Guha R, Shinn P, Young RM, Keller JM, Liu D, Goldlust IS, Yasgar A, McKnight C, Boxer MB, et al. (2014). High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell–like diffuse large B-cell lymphoma cells. Proc. Natl. Acad. Sci 111, 2349–2354. [PubMed: 24469833]

Harding SD, Sharman JL, Faccenda E, Southan C, Pawson AJ, Ireland S, Gray AJG, Bruce L, Alexander SPI, Anderton S, et al. (2018). The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Res 46, D1091–D1106. [PubMed: 29149325]

Hopkins AL, and Groom CR (2002). The druggable genome. Nat. Rev. Drug Discov 1, 727. [PubMed: 12209152]

Inglese J, Auld DS, Jadhav A, Johnson RL, Simeonov A, Yasgar A, Zheng W, and Austin CP (2006). Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc. Natl. Acad. Sci. U. S. A 103, 11473–11478. [PubMed: 16864780]

Jones DS, Jenney AP, Swantek JL, Burke JM, Lauffenburger DA, and Sorger PK (2016). Profiling drugs for rheumatoid arthritis that inhibit synovial fibroblast activation. Nat. Chem. Biol 13, nchembio.2211.

Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT, et al. (2008). A quantitative analysis of kinase inhibitor selectivity. Nat. Biotechnol 26, 127. [PubMed: 18183025]

Kazandjian D, Blumenthal GM, Chen H-Y, He K, Patel M, Justice R, Keegan P, and Pazdur R (2014). FDA Approval Summary: Crizotinib for the Treatment of Metastatic Non-Small Cell Lung Cancer With Anaplastic Lymphoma Kinase Rearrangements. The Oncologist 19, e5–e11. [PubMed: 25170012]

Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, Wang Z, Dohlman AB, Silverstein MC, Lachmann A, et al. (2017). The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst Khalil HS, Mitev V, Vlaykova T, Cavicchi L, and Zhelev N (2015). Discovery and development of Seliciclib. How systems biology approaches can lead to better drug performance. J. Biotechnol 202, 40–49. [PubMed: 25747275]
Kimura S, Naito H, Segawa H, Karoda J, Yuasa T, Sato K, Yokota A, Kamitsuji Y, Kawata E, Ashihara E, et al. (2005). NS-187, a potent and selective dual Bcr-Abl/Lyn tyrosine kinase inhibitor, is a novel agent for imatinib-resistant leukemia. Blood 106, 3948–3954. [PubMed: 16105974]

Klaeger S, Heinzlmier S, Wilhelm M, Polzer H, Vick B, Koenig P-A, Reinecke M, Ruprecht B, Petzoldt S, Meng C, et al. (2017). The target landscape of clinical kinase drugs. Science 358, eaan4368.

Kryštof V, Lenobel R, Havlíček L, Kuzma M, and Strnad M (2002). Synthesis and biological activity of olomoucine II. Bioorg. Med. Chem. Lett 12, 3283–3286. [PubMed: 1239273]

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313, 1929–1935. [PubMed: 17008526]

Lemieux GA, Liu J, Ashrafi K, Mayer N, Bainton RJ, and Werb Z (2011). A whole-organism screen identifies new regulators of fat storage. Nat. Chem. Biol 7, nchembio.534.

Marks PA (2007). Discovery and development of SAHA as an anticancer agent. Oncogene 26, 1210204.

Moffat JG, Rudolph J, and Bailey D (2014). Phenotypic screening in cancer drug discovery — past, present and future. Nat. Rev. Drug Discov 13, nrd4366.

Niepel M, Hafner M, Pace EA, Chung M, Chai DH, Zhou L, Schoeberl B, and Sorger PK (2013). Profiles of Basal and Stimulated Receptor Signaling Networks Predict Drug Response in Breast Cancer Lines. Sci. Signal 6.

Niepel M, Hafner M, Duan Q, Wang Z, Paul EO, Chung M, Lu X, Stuart JM, Golub TR, Subramanian A, et al. (2017). Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nat. Commun 8, 1186. [PubMed: 29084964]

Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Guha R, Hersey A, Holmes J, et al. (2018). Unexplored therapeutic opportunities in the human genome. Nat. Rev. Drug Discov 17, 317–332. [PubMed: 29472638]

Oumata N, Bettayeb K, Ferandin Y, Demange L, Lopez-Giral, A, Goddard M-L, Myrianthopoulos V, Mikros E, Flajolet M, Greengard P, et al. (2008). Roscovitine-Derived, Dual-Specificity Inhibitors of Cyclin-Dependent Kinases and Casein Kinases 1. J. Med. Chem 51, 5229–5242. [PubMed: 18698753]

Poon CC, and Kelly JJ (2017). Development of crizotinib, a rationally designed tyrosine kinase inhibitor for non-small cell lung cancer. Int. J. Cancer 140, 1945–1954. [PubMed: 27874172]

Roehrborn CG, Spann ME, Myers SL, Serviss CR, Hu L, and Jin Y (2015). Estrogen receptor beta agonist LY500307 fails to improve symptoms in men with enlarged prostate secondary to benign prostatic hypertrophy. Prostate Cancer Prostatic Dis 18, 43–48. [PubMed: 25348255]

Rogers D, and Hahn M (2010). Extended-Connectivity Fingerprints. J. Chem. Inf. Model 50, 742–754. [PubMed: 20426451]

Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, and Kuriyan J (2000). Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science 289, 1938–1942. [PubMed: 10988075]

Shah NP, Tran C, Lee FY, Chen P, Norris D, and Sawyers CL (2004). Overriding Imatinib Resistance with a Novel ABL Kinase Inhibitor. Science 305, 399–401. [PubMed: 15256671]

Stribl SM (2014). Overcoming Drug Development Bottlenecks With Repurposing: Old drugs learn new tricks. Nat. Med 20, 590–591. [PubMed: 24901567]

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 171, 1437–1452.e17. [PubMed: 29195078]

Swinnen DC, and Anthony J (2011). How were new medicines discovered? Nat. Rev. Drug Discov 10, 507. [PubMed: 21701501]

Wang J, and Gray NS (2015a). SnapShot: Kinase Inhibitors I. Mol. Cell 58, 708e1. [PubMed: 26000854]

Wang J, and Gray NS (2015b). SnapShot: Kinase Inhibitors II. Mol. Cell 58, 710e1. [PubMed: 26000855]
Wang Y, Cornett A, King FJ, Mao Y, Nigsch F, Paris CG, McAllister G, and Jenkins JL (2016). Evidence-Based and Quantitative Prioritization of Tool Compounds in Phenotypic Drug Discovery. Cell Chem. Biol 23, 862–874. [PubMed: 27427232]

Wassermann AM, Camargo LM, and Auld DS (2014). Composition and applications of focus libraries to phenotypic assays. Front. Pharmacol 5.

Workman P, and Collins I (2010). Probing the Probes: Fitness Factors For Small Molecule Tools. Chem. Biol 17, 561–577. [PubMed: 20609406]

Yeo W, Goh B, Le Tourneau C, Green SR, Chiao JH, and Siu LL (2009). A phase II randomized study of oral seliciclib in patients with previously treated nasopharyngeal carcinoma. J. Clin. Oncol 27, 6026–6026.
SIGNIFICANCE

Libraries of small molecules are frequently used to perform chemical genetics, drug discovery and therapeutic repurposing. Many libraries have been made available, however, there are few data-driven methods to compare different libraries or create new ones. Major challenges in creating data-driven compound collections is the different data-types through which compounds’ performances can be evaluated as well as the vast incompleteness of data. Here, we describe tools to compare compounds and libraries on binding selectivity, target coverage, induced cellular phenotypes, chemical structure and phase of clinical development – while appropriately addressing data incompleteness. In general, our approach aims to include several compounds with lowest possible overlap of off-targets per target of interest. We furthermore create web-based tools to select chemical probes and to compose small molecule libraries, available via http://www.smallmoleculesuite.org, where we allow additional customization of libraries based on user preferences. Using our approach, we develop two new small molecule libraries; LSP-OptimalKinase and LSP-MoA, which target the kinome and the liganded genome, specifically. We envision our tools to be of assistance to molecular biology, chem- and bioinformatics, machine learning, drug discovery and drug repurposing.
Highlights

- Existing small molecule collections vary greatly on selectivity and target coverage
- A data-driven approach to library design enhances diversity and library performance
- The *LSP-OptimalKinase* library enhances selectivity and coverage for kinome targets
- The *LSP-MoA* library optimally targets 1852 genes in the liganded genome
Figure 1. Structural comparisons across the libraries.
(A) Number of identical compounds in six widely used libraries. Compounds were matched based on chemical structure; Dundee denotes a compound assembled by the University of Dundee; SP the SelleckChem Pfizer licensed collection; SK denotes the SelleckChem kinase collection; EMD a kinase inhibitor collection sold by Tocris Bioscience; PKIS the “Published Kinase Inhibitor Set” assembled by Glaxo; and LINCS, the HMS-LINCS small molecule collection (see Figure S1, Table S1 for references and Web links). Asterisk highlights library files provided by EMD that did not contain the same number of compounds as advertised online. (B) Structural similarity for LINCS and PKIS libraries calculated using the Tanimoto similarity of Morgan fingerprints (at 2Å); molecules in two representative clusters are shown below. Data on these compounds could not be retrieved through customer service; thus only structures present in the available library files are
considered in this paper. (C) The distribution of compound clusters for different libraries (with a similarity cut-off of 0.7).
Figure 2. Comparison of induced cellular phenotypes across kinase inhibitor libraries.
(A) Left panel: phenotypic fingerprint for imatinib (yellow) and Nilotinib (green), which have a Pearson correlation of 0.82. Right panel: phenotypic fingerprint for imatinib (yellow) and dasatinib (red), which have a Pearson correlation of 0.16. Only assays for which either compound exhibited a robust z-score (r-score) greater than 3 or smaller than −3 are shown.
(B) Distribution of pairwise correlations among phenotypic fingerprints across compounds in different libraries normalized to the number of compounds for which phenotypic information is available in ChEMBL.
(C) Comparison of ChEMBL-derived PFP and L1000 Connectivity Map (C-MAP) correlations. Black and red dots highlight compounds targeting the PI3K-AKT-MTOR or RAF-MEK-ERK pathways (see text for details).
(D) Number of compounds in various phases of clinical development for each library.
Figure 3. Comparisons of selectivity across libraries.
(A) Illustrative example of selectivity assertions with incomplete target information. Estimated binding affinities for dasatinib (right) and PP121 (left) are shown on a kinase tree. Red circles denote confirmed binding (bigger circles indicate higher affinity), black squares denote confirmed non-binding ($K_d > 10\,\text{uM}$), the absence of data point (no color) denotes an absence of information. (B) Features of compound selectivity classes. Note that poly-selective (PS) and unknown (UN) selectivity classes require that first quartile of all affinities measured be less than 9000 nM. (C) Pie charts show the number of kinases inhibited within a specified selectivity class per library. The highest selectivity achieved for each of the 545 human kinases is shown.
Figure 4. Comparisons of target affinity spectra across libraries.

(A) Schematic illustrating the conversion of three types of data on target binding (dose response, one-point inhibition and manual annotation) into the binding assertions summarized by TAS vectors. Assertions of 1 (yellow) correspond to confirmed high-affinity binding. Assertions of 10 (red) correspond to confirmed non-binding. Target assertions were used to quantify similarities in target affinity spectra. (B) Distribution of target affinity similarities (calculated using weighted Jaccard similarities) for compounds having the same (blue) or different (brown) nominal targets. Inset table depicts the number of compounds used for the analysis. (D) Illustrative examples of the lack of correlation between a target spectrum and nominal drug targets. PD-173074 and Cabozantinib illustrate a pair of compounds with the very similar nominal targets but different target spectra. BMS-777607 and ponatinib illustrate a pair with different nominal targets, but similar target spectra. Black arrows denote nominal target(s); targets are listed in alphabetical order by affinity class. As
it happens, in the example shown neither compound had a target binding weight corresponding to an estimated $K_d$ of 1–10 µM. See also Figure S2-S5.
Figure 5. Comparison of structural, phenotypic and target similarity.
(A) Comparison of phenotypic correlation and structural similarity for all libraries. Only a very small subset of the data is correlated. Dotted line denotes the threshold of 0.2 used for library design. (B) Conditional p-values for phenotypic-correlations given a specified minimal degree of structural similarity. (C) Comparison of phenotypic correlation and structural similarity for all Kinase inhibitors in ChEMBL. All inhibitors bind at least one kinase with Kd ≤ 100 nM. (D) Comparison of phenotypic correlation and structural similarity for all Kinase inhibitors in ChEMBL. All inhibitors bind at least one GPCR with Kd ≤ 100 nM. (E) Comparison of target similarity and structural similarity for the LINCS library and (F) for the PKIS library. In panel C, compound-pairs with same nominal target are plotted in blue, compound-pairs with different nominal targets are plotted in brown. The marginal distributions for all compound pairs are plotted in magenta on the axes.
Figure 6. Creating a custom chemical genetics library.

(A) Illustration of the compound selection procedure for a query target “BRAF” based on selectivity score and clinical phase. In this case, all compounds shown belong to Tier A as they either have MS selectivity or are clinically approved drugs. (B) Overview of the tiered design of the MoA and LSP-OptimalKinase libraries based on selectivity class and phase of clinical development. (C) Number of kinases inhibited at a specified selectivity class based for the LSP-OptimalKinase Library. The highest selectivity achieved for each of the 545 human kinases is shown. (D) Clinical grade compounds in the LSP-Optimal Kinase library by tier. (E) Target coverage by selectivity class for the liganded in the LSP-MoA library. (F) Selectivity distribution per target class in LSP-MoA.

“Full” full refers to the complete target class and “dark” to the understudied portion of the class as curated by the NIH IDG program (see text for details and Table S4). See also Figure S4, S6.