Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections

Paul A. Clemons1,2, J. Anthony Wilson3, Vladó Dančík4, Sandrine Muller4, Hyman A. Carrin5ki, Bridget K. Wagner4, Angela N. Koehler1, and Stuart L. Schreiber2,3,6,7

1Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142; 2Howard Hughes Medical Institute, 7 Cambridge Center, Cambridge, MA 02142; and 3Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Using a diverse collection of small molecules we recently found that compound sets from different sources (commercial; academic; natural) have different protein-binding behaviors, and these behaviors correlate with trends in stereochemical complexity for these compound sets. These results lend insight into structural features that synthetic chemists might target when synthesizing screening collections for biological discovery. We report extensive characterization of structural properties and diversity of biological performance for these compounds and expand comparative analyses to include physicochemical properties and three-dimensional shapes of predicted conformers. The results highlight additional similarities and differences between the sets, but also the dependence of such comparisons on the choice of molecular descriptors. Using a protein-binding dataset, we introduce an information-theoretic measure to assess diversity of performance with a constraint on specificity. Rather than relying on finding individual active compounds, this measure allows rational judgment of compound subsets as groups. We also apply this measure to publicly available data from ChemBank for the same compound sets across a diverse group of functional assays. We find that performance diversity of compound sets is relatively stable across a range of property values as judged by this measure, both in protein-binding studies and functional assays. Because building screening collections with improved performance depends on efficient use of synthetic organic chemistry resources, these studies illustrate an important quantitative framework to help prioritize choices made in building such collections.

A central theme in applying cheminformatics to discovery chemistry is to relate synthetic decisions to consequences on both chemical structure and biological assay performance. Historically, such efforts focused on small sets of similar compounds, and single performance measurements (1–3), providing guidance to chemists in compound optimization against single-target proteins or processes (4). However, additional methods are needed to judge large sets of compounds, such as those used in small-molecule screening. Progress toward more valuable screening collections (5) requires unbiased methods to evaluate diversity of assay performance for compound sets rather than performance of individual members.

A widely used method to judge compounds for drug discovery is the “rule of 5” (ROS) (6), which predicts poor absorption or permeation for compounds that deviate from property-value constraints: H-bond donors (Hd) and acceptors (Ha), molecular weight (MW), and calculated partition coefficients (cLogP). Recent studies have attempted to refine such rules (7–9) and extend them to other goals (10–13), such as making leads or probes. Such property filters have been debated and reviewed (14–16), and their long-term impact on pharmaceutical research is starting to be analyzed (17, 18). Importantly, exceptions to these rules, including natural products (19–21), are well-noted and suggest that previously undescribed types of chemistry might access property distributions acceptable for certain goals despite nonadherence to established rules.

Comparative analyses of compound sets usually use computed properties (19, 22, 23) or historical assay results (24, 25). Significant progress has been made quantifying and visualizing properties of compound sets (26), including methods that relate structure to intuitive notions of shape (27–29), and similarity fusion methods (30–33) that describe relationships between sets. Moreover, chemical similarity and diversity analyses continue to progress (34–37), including studies using Shannon entropy (38) as a measure of structure information among compounds (39–41), addressing reagent selection (42), database similarity searches (43), and scaffold diversity (44). Entropy-based methods have also been used on assay data to distinguish single-target compounds from those with multitarget effects (45), and to quantify relationships between targets based on Ki profiles among sets of common inhibitors (46).

Despite advances in cheminformatics, methods to measure assay performance of compound collections remain underexplored. One important study focused on compounds from different sources, including drugs (19). Other studies focused on molecular complexity, suggesting intermediate complexity is preferable for drug leads (17, 47). Recently, we investigated relationships between intermediate stereochemical complexity and binding specificity (48). What these previous studies did not address is set-based behavior of compound collections. In screening collections, the value of chemistry investment needs to be measured in terms of overall collection performance, rather than anecdotes about the best performers. An assumption often made is that diverse structures will result in diverse outcomes across many assays, but few studies address this question directly (49–53). Likewise, our recent analysis (48) did not account for the distinction between individual and groupwise compound performance. The availability of data from large-scale experiments (48) and public databases provides an opportunity to measure set-based performance quantitatively, rather than measuring success by finding a few “special” compounds.

We recently analyzed a large compound collection in 100 parallel protein-binding assays (48) and found both protein-binding frequencies and specificities are increased among compounds having intermediate stereochemical complexity. Here, we present a more general method for evaluating diversity in assay performance for compound sets as a measure of structure information among compounds (39–41), addressing reagent selection (42), database similarity searches (43), and scaffold diversity (44). Entropy-based methods have also been used on assay data to distinguish single-target compounds from those with multitarget effects (45), and to quantify relationships between targets based on Ki profiles among sets of common inhibitors (46).

Author contributions: P.A.C., B.K.W., A.N.K., and S.L.S. designed research; P.A.C., J.A.W., V.D., and A.N.K. performed research; P.A.C., J.A.W., V.D., S.M., H.A.C., B.K.W., and A.N.K. contributed new reagents/analytic tools; P.A.C., S.M., H.A.C., and A.N.K. analyzed data; and P.A.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. E-mail: pclemons@broadinstitute.org.

2On leave from: Mathematical Institute, Slovak Academy of Sciences, Košice, 040 01, Slovakia.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1015024108/-/DCSupplemental.
we extend analysis of the same compounds to physicochemical properties (6, 7, 16, 54) and shape-based descriptors (27, 29), revealing additional similarities and differences between sets. We use principal component analysis (PCA) (55) of chemically intuitive properties to analyze sources of variation among the sets. Different descriptors report different aspects of chemical structure (36), and we aim to illustrate for synthetic chemists how quantitative relationships can be reconciled with chemical intuition. We also provide a framework to evaluate performance diversity of compound sets using Shannon entropy (38) on profiles of assay measurements. In this context, entropy measures how evenly a compound set is distributed over all possible patterns of performance accessible from a given set of measurements. We apply this method to both protein-binding profiles (48) and diverse profiles of functional assays extracted from ChemBank (56). These methods provide a quantitative measure for performance evaluation of small-molecule collections (5, 48) and encourage exploration of emerging relationships between performance diversity and molecular property distributions.

**Results**

We characterized compounds from three sources that were exposed to 100 parallel protein-binding assays (48). The compound collection consists of (i) 6,152 compounds representative of screening collections (commercial compounds; CC); (ii) 2,477 naturally occurring compounds (natural products; NP); and (iii) 5,963 compounds from the academic synthetic chemistry community (diverse compounds; DC’). These sets provide an opportunity to compare properties and performance of compounds from different origins (cf. ref. 19), including one group (DC’) whose property distributions differ from other compounds, and whose properties and performance have not thoroughly been investigated.

Not surprisingly, when applying established filtering criteria (16), nearly all CC will “pass” ROS (99.9%) (6, 16) or an alternative based on polar surface area (PSA) and rotatable bonds (Rot) (99.7%) (7, 16). In contrast, up to approximately 1/3 of DC’ (73.0% ROS, 66.5% alternative) and up to approximately 2/5 of NP (71.1% ROS, 60.7% alternative) would “fail.” Because differences between natural products and typical screening compounds are established (16, 19, 21, 57), we sought to compare DC’ with each of CC and NP to determine which set DC’ more resembles for each of six common properties (16). For five of the six properties, DC’ is more similarly distributed to either CC or NP than the latter two are to each other, and for three of these, the values for DC’ are intermediate between those of CC and NP (Table 1). That DC’ is heavier and more lipophilic than CC or NP is evident by inspecting the Table 1 and visualizing MW versus cLogP (Fig. 1A). In a PCA “chemical space” composed of these six properties (Fig. 1B), both NP and DC’ overlap substantially with the more compact CC, but DC’ is more similar to NP along one dimension (Fig. 1C) and more different in another (Fig. 1D). These results show that some members of DC’ access part of the space not accessed by either CC or NP, suggesting properties and performance for DC’ should be evaluated in their own right, rather than being presumed similar to either CC or NP.

To refine our chemical intuition about similarities and differences among CC, NP, and DC’, we analyzed PCA coefficients to learn which properties are correlated (Fig. 2A), and how these properties vary in each set (Fig. 2B). Not surprisingly, PSA is correlated with Ha and Hd, and NP has high variation in this direction, consistent with other studies (19, 21). Similarly, Rot is correlated with MW, and these properties, along with cLogP, account for much of the variation in DC’. Previously, we classified these compounds for binding specificity in protein-binding profiles (48). Using similarly defined specificity groups (“specific”: bound 1 protein; “intermediate”: bound 2–5 proteins; “promiscuous”: bound 6+ proteins), we asked whether compounds in different groups are concentrated or distributed in the property space (Fig. 2C). We found that specific and intermediate compounds, regardless of source, are well-distributed throughout the space, whereas promiscuous compounds are significantly concentrated in the center (p < 0.0099). Because the center of the space corresponds to common property filters, this result suggests that our binding experiments accessed a greater number of desirable outcomes (specific binding) than had we restricted ourselves to compounds passing common filters.

Although the above properties allow rapid application of filters, they tend to oversimplify relationships between compounds and therefore offer only partial guidance to chemists planning syntheses. To illustrate additional possibilities for guidance, we characterized relationships among CC, NP, and DC’, using three chemical spaces that shed light on different aspects of structure variation (Fig. 3). First, we considered atom counts: PCA of this chemical space reveals that carbon, nitrogen, and oxygen are the dominant contributors to variation (Fig. 3A). Notably, NP achieves most of its variation with oxygen and carbon, with little variation in nitrogen composition (Fig. 3B). In contrast, both CC and DC’ get most compositional variation from nitrogen and carbon. Second, to examine compounds in terms of ring and side-chain content, we used descriptors counting rings, side chains, and branches (Fig. 3C and D). Here, CC varies primarily in number of aromatic rings, as does DC’ (with higher variation in total rings). In contrast, NP has high variation in side-chain number (unconnected fragments after removing ring atoms), but less variation in number of aromatic rings. Third, to extend our previous analysis (48) of electronic character of carbon atoms, we used electropotential-chemical (E-state) descriptors for nine carbon environments (Fig. 3E). These descriptors measure variation in electronic environment (54) for these carbon types, rather than

**Table 1. Distributions of properties typically used for filtering (e.g., Lipinski ROS)**

<table>
<thead>
<tr>
<th>Property</th>
<th>CC Median</th>
<th>CC Mean</th>
<th>CC SD</th>
<th>NP Median</th>
<th>NP Mean</th>
<th>NP SD</th>
<th>DC’ Median</th>
<th>DC’ Mean</th>
<th>DC’ SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular weight (Da)</td>
<td>311</td>
<td>314</td>
<td>75</td>
<td>311</td>
<td>314</td>
<td>75</td>
<td>311</td>
<td>314</td>
<td>75</td>
</tr>
<tr>
<td>Calculated logP</td>
<td>3.2</td>
<td>3.2</td>
<td>1.6</td>
<td>68</td>
<td>72</td>
<td>31</td>
<td>95</td>
<td>100</td>
<td>41</td>
</tr>
<tr>
<td>Polar surface area (Å²)</td>
<td>104</td>
<td>136</td>
<td>93</td>
<td>96</td>
<td>100</td>
<td>41</td>
<td>96</td>
<td>100</td>
<td>41</td>
</tr>
</tbody>
</table>

Bold pairs of values for each property indicate more similarly distributed pairs of sets. Also indicated is whether the distribution for DC’ is centered between (underline) or outside (italics) those of CC and NP.

**Fig. 1.** DC’ occupies a distinct part of the property space from CC and NP. (A) Scatterplot of MW versus cLogP, omitting three compounds (all from NP) with MW > 1,500 and three compounds (all from DC’) with cLogP > 15. (B) Top three principal components (PC) (93% of total variance) using six properties. (C) PC1 versus PC3, illustrating dimension in which DC’ is similar to NP. (D) PC2 versus PC3, illustrating dimension in which DC’ is distinct from NP (CC: dark red; NP: dark green; DC’: dark blue).
Their variation around less-connected sp2 carbon atoms connected to three or four other heavy atoms (ssssC, ssCH3). In contrast, CC and DC′ express more of their variation around less-connected sp3 carbons (CH3, CH2) or sp2 carbons (aaCH, aac). Often, descriptors such as E states correspond to variation in electronic character. Recently, we used principal moment-of-inertia (PMI) descriptors (27) of 3D shape to analyze chemist decisions during diversity-oriented syntheses, including relationships between skeletal and stereochemical diversity (60, 61) and a unique fragment library (62). PMI descriptors provide an intuitive notion of molecular shape, occupying a triangular map bounded by canonical shapes (rod, disk, and sphere). We mapped the three sets (CC, NP, and DC) and the specificity groups (promiscuous, intermediate, and specific) to PMI plots (Fig. 4A). Consistent with earlier observations, CC is dominated by flat and linear compounds. NP covers more of the PMI map, its distribution centered up and to the right relative to CC. Notably, NP has less coverage in the disk-like (bottom center) region of the map than either CC or DC′. DC′ covers more of the shape space than NP, with its distribution centered down and to the right of CC. DC′ also contains low density on the extreme left diagonal of the space, indicating few very flat or linear compounds. The specificity groups also follow a trend, with less difference between them than the compound sets.

To quantify shifts in PMI density, we measured distance distributions for each set or group relative to vertices of PMI space. Cumulative distance distributions allow us to quantify the statistical significance of these differences, with all three of CC, NP, and DC′ significantly different relative to the sphere shape (Fig. 4B), and DC′ less rod-like and more disk-like than either CC or NP. Importantly, more specific binders (regardless of set) are significantly more sphere-like than promiscuous ones. To refine this analysis, we also computed distance distributions from canonical flat and spherical shapes based on alpha-shaped descriptors (29). The results for sphere likeness exactly match those with PMI distance to the sphere vertex, and for flat shapes significantly discriminate CC, NP, and DC′. Moreover, promiscuous compounds are significantly flatter in their distance distribution (Fig. 4C).

The foregoing analyses treat properties as distributions, but until now we have considered assay performance for each compound separately, for example, assigning each to a specificity group. An important consideration for compound collections is how well they access a diversity of assay outcomes. Given 100 specific compounds (each binds exactly one of 100 proteins), compare the case where each compound binds the same protein to the case where each compound binds a different protein: “Hit” rates and specificity groups are not sufficient. To evaluate compound sets, we need to distinguish these cases. Shannon entropy (38) and related measures (45) provide a framework to do so when applied to matrices of assay outcomes (cf. figure 4 in ref. 48). Entropy applied to profiles (profile entropy) of small-molecule performance allows us to distinguish the possibilities above, providing a higher score for more diverse sets of assay outcomes, a maximal score for uniform coverage of all possible outcomes, and penalties for missing outcomes or “dilution” with inactive compounds. By itself, however, profile entropy does not distinguish profiles based on their selectivity. A profile with a single assay activity
is equivalent to one with a single assay inactivity (i.e., a promiscuous compound). To address this concern we define a weighted measure (weighted profile entropy) to reward more specific profiles over less specific ones.

Using profile entropy, we analyzed the performance of CC, NP and DC′ using profiles of binding to 100 proteins (Fig. 5A–D) (48). By this measure, CC exhibits higher performance diversity than DC′, and substantially higher than NP, possibly due to lower hit rates in NP (48). However, considering hit compounds only (bound at least one protein) does not change relative scores; NP and to a lesser extent DC′, “concentrates” its compounds on particular patterns of protein binding (which entropy penalizes). Weighted profile entropies show similar relationships between sets, as expected because most hit compounds bind one (65.3%) or two (13.2%) proteins. As an independent test of performance diversity for the compound sets, we extracted functional assay data from ChemBank corresponding to all 14,592 compounds. Each compound had been exposed to between 51 and 154 different functional assay readouts (median = 71, mean = 85). Using these data, we computed profile entropies for CC, NP, and DC′ (Fig. 5B). The functional assay results exhibit higher overall entropies (a smaller fraction is never called active; 30.4% versus 78.4%), but performance diversity remains highest for CC and lowest for NP, except in weighted profile entropies, where DC′ drops below NP, suggesting that either nonspecific effects or correlations between assay readouts are more common for DC′ than for NP.

Finally, we revisited the idea that computed property distributions will impact compound set performance: specifically the ability to produce diverse activities in primary assay formats (binding or functional). We computed profile entropies for subsets of compounds with similar property values. We took sets of profiles of equal size, centered on a compound with the (ranked) property value of interest. We found that increasing cLogP increased performance diversity in protein-binding profiles using both unweighted and weighted profile entropy measures (Fig. 5C). In contrast, though increasing both MW and cLogP produced increased performance diversity in unweighted functional assay profiles, both properties showed stable performance diversity when specificity-weighted profile entropies were considered (Fig. 5D), suggesting increases in the unweighted case may stem from nonspecific effects in assays. Similar observations were observed with several descriptors (see SI Datasets D1–D8); for example, we observed that increasing PSA to extreme values decreased performance diversity in functional assays, consistent with the inability of compounds to penetrate cells. Clearly, other such comparisons are warranted, but these preliminary results illustrate a framework to evaluate set-based performance in terms of calculated property distributions.

**Discussion**

Quantifying properties and behavior of compound sets has implications for organic synthesis toward small-molecule probes and drugs. In particular, it is valuable to make quantitative decisions about choices: which compounds to buy, which to synthesize, which to include in a screening collection, which to retire from screening. Decisions regarding individual compounds typically dominate such discussions: Individual compounds pass or fail filtering rules, individual compounds hit or not in assays, etc. In this study, we use real-world compounds and data to illustrate methods that can guide decisions about sets of compounds.

Our results reflect the properties a particular set of compounds (48), but the message of this study is *not a statement about these particular compounds*. Indeed, our current discovery collection at the Broad Institute reflects many considerations (26, 63–67) that were never applied to the compounds in this study. Rather, we aim to promote a way of thinking about compound sets that is unbiased, quantitative, and decision-oriented, not to prescribe which particular decisions should be made. Depending on the goal of any particular study, attention to different properties or performance is warranted. For example, Figs. 1 and 2 remind us that although for certain applications (e.g., drug discovery) additional property constraints might be applied, we should be cautious against overapplication of constraints that might obscure previously undescribed insights into small-molecule behavior—lack of “drug-likeness” does not mean that a compound’s profile
of interactions with 100 different proteins is not valuable information.

Compounds are often judged by whether they add “diversity” to a compound collection. Fig. 3 reminds us that diversity is a relative concept, depending on the sets being compared and properties used to describe them. Consistent with earlier studies (60, 61), visualizing molecular shape relative to canonical shapes (27–29) provides a powerful stimulus to chemists interested in “escaping flatland” (68) by making more “globular” compounds. Importantly, we show in Fig. 4 how these relationships can be quantified and their significance assessed, but our results do not support a causal relationship between globularity and specificity. DC’ is more globular (sphere-like) than other sets, and the group of specific binders is more globular than promiscuous compounds. However, this is primarily because DC’ is enriched in specific binders (48), not because globularity “imparts” specificity. Specific members of CC and NP are not statistically enriched for globularity relative to their promiscuous counterparts, suggesting more complicated relationships that warrant investigation with different compounds.

An important advance in this study is extending Shannon entropy (38) to profiles of assay performance for compound sets. Entropy has been used extensively for chemical structure analysis (39–44), but has been underutilized in the area of measuring diversity of assay performance (45, 46). Here, we apply entropy broadly to multiasay performance profiles (51, 53) for large collections and observe that CC accesses 654 unique protein-binding profiles with 1,406 active compounds, NP accesses 112 profiles with 324 compounds, and DC’ accesses 422 profiles with 1,406 compounds. Entropy is sensitive precisely to these relationships—how uniformly are compounds distributed over different patterns of performance? Weighted profile entropy addresses an important concern about specificity among binding profiles, and future studies should reconcile weighting with further analysis of on-target and off-target effects in phenotypic assays and provide improved handling of missing data (45). As suggested by Fig. 5, profile entropy can be sensitive to the size of compound sets and the number of assays. Moreover, our nonparametric analysis of entropy relative to property distributions is likely sensitive to the shapes of these distributions. Additional studies with other datasets are needed to refine the methodology, but the speed of profile entropy calculations, and their generality to any small-molecule profiling study (69), are indicators of future promise.

Overall, this study focuses attention on the need of computational methods to adapt to the changing requirements of high-throughput chemistry and screening. It provides a framework for analysis of compound collections that focuses on overall collection performance rather than performance of individual members (as with more conventional structure-activity studies). We intend this study as a follow-up to our prior work (48), in that it provides informative additional analysis of compounds from different sources. However, we also intend this work to advance methods for analysis that can be applied to previously undescribed compounds and profiling datasets in the future. Such large-scale computational analysis of compound sets in the context of screening data can influence synthetic (or acquisition) decisions leading toward the assembly of improved screening collections.

**Materials and Methods**

Chemical structures, categories, sources, and ChemBank (56) identifiers for each compound are already published (48), and we used this information as the source of structures and binary protein-binding data. We calculated molecular property counts (SI Datasets D1 and D2) and E-state descriptors using Pipeline Pilot (Accelrys, Inc.). We also considered these descriptors as a function of protein-binding specificity (SI Dataset D3). For 3D descriptors, we used a ChemAxon module employing DREIDING force field (70) in Pipeline Pilot to generate up to 16 3D conformers per molecule, retaining those with 3 kcal/mol of the lowest energy conformer. We computed PMI (27) and alpha-shape (29) descriptors using MATLAB code (The MathWorks, Inc.) that follows published methods. Median PMI values and alpha-shape-based distances were taken across retained conformers. In PCA visualizations, small numbers of outliers are not shown for graphical clarity, but all compounds were included in calculations. Significance of spread or concentration in PCA was performed between distributions of Hotelling’s T² values, and in PMI and alpha-shape plots between distance distributions (SI Dataset D4), each using Kolmogorov-Smirnov tests (71). Correlations between 2D and PMI descriptors are also provided (SI Dataset D5).

For each compound we constructed a performance profile assigning binary (0,1), binding data) or discrete (−1,0,1), functional data) values representing activity for a compound in a given assay, collecting such values across all assays into a vector x. Discretization of binding data are described elsewhere (48), and ChemBank data were handled similarly, except both high- and low-signal outlier values (56) were accepted (SI Datasets D6 and D7). All distinct performance profile vectors x for compounds were collected to set S, and Shannon entropy (H) was computed by calculating relative frequencies p(x) and summing frequency terms over x ∈ S: H(S) = −Σ(x) log₂(p(x)) (profile entropy). To calculate entropy for a set of profiles weighted by a specificity constraint, we first computed H separately for each subset Sᵢ of profiles sharing the same number m of nonzero profile features, then computed a weighted sum of these entropies with weighting factor wᵢ = exp(−ln(2) mᵢ), to give Hₛ(S) = ΣwᵢHₛᵢ(S) (weighted profile entropy). For ChemBank profiles, missing data were censored to zero before entropy calculations (SI Dataset D8). Statistical analyses, visualizations, and entropy calculations were performed in MATLAB.

Clemons et al.


