Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles


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

Contributed by Stuart L. Schreiber, September 3, 2010 (sent for review August 9, 2010)

Using a diverse collection of small molecules generated from a variety of sources, we measured protein-binding activities of each individual compound against each of 100 diverse (sequence-unrelated) proteins using small-molecule microarrays. We also analyzed structural features, including complexity, of these small molecules. We found that compounds from different sources (commercial, academic, natural) have different protein-binding behaviors and that these behaviors correlate with general trends in stereochemical and shape descriptors for these compound collections. Increasing the content of sp3-hybridized and stereogenic atoms relative to compounds from commercial sources, which comprise the majority of current screening collections, improved binding selectivity and frequency. The results suggest structural features that synthetic chemists can target when synthesizing screening collections for biological discovery. Because binding proteins selectively can be chemists can target when synthesizing screening collections for frequency. The results suggest structural features that synthetic chemists can target when synthesizing screening collections for biological discovery. Because binding proteins selectively can be.

Results

To quantify stereochemical and shape complexity, we calculated two parameters for each compound for comparative analysis between the three compound sources. While it is likely that more complex descriptors may better physically model molecular shape complexity, one motivation of the current study was to link simple size-independent metrics with compound performance. First, we defined stereochemical complexity as the proportion of carbon atoms that are stereogenic (Cstereogenic/Ctotal). This metric provides a size-independent global assessment of stereochemical complexity, varying on the range [0,1] for each molecule. Inspecting histograms of this metric as a function of compound source (Fig. 2) revealed that CC is lowest in stereochemical complexity (median = 0.00; mean = 0.022). In contrast, NP is highest in stereochemical complexity (median = 0.24; mean = 0.24), while DC has intermediate values (median = 0.11; mean = 0.12).

Second, we defined shape complexity as the ratio of sp3-hybridized carbon atoms to total sp3- and sp2-hybridized carbons (Csp3/(Csp2 + Csp3)). This metric is similar to the recently reported Fsp3 metric (14); again, this metric is size-independent, varying on the range [0,1] for each molecule. Using this metric (Fig. 3A), we observed that CC is lowest in proportion of sp3-hybridized carbons (median = 0.22; mean = 0.27), NP is highest (median = 0.55; mean = 0.55), and DC has an intermediate distribution (median = 0.36; mean = 0.39). When we restricted our analysis to carbon atoms in the molecular scaffold (15), we observed a decrease in overall sp3 carbon proportions in all three populations (Fig. 3B). This effect was most striking in CC (median = 0.071; mean = 0.16), indicating that a substantial portion of sp3 carbons in these molecules are in appendages rather than in (predominantly flat) core skeletons. In contrast, NP molecules retain a large proportion of their sp3 carbons in core skeletons (median = 0.50; mean = 0.49).


The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

To whom correspondence may be addressed. E-mail: pclemons@broadinstitute.org or stuart_schreiber@harvard.edu.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1012741107/-/DCSupplemental.
In addition to analyzing computed properties, we sought to determine differences in protein-binding abilities of compounds from different origins. For this study, we analyzed a dataset resulting from testing all members of the compound collection individually in binding assays against each of 100 different purified proteins using the SMM platform, which gives a preliminary indication of protein binding. In terms of sequence, proteins were selected having a wide range of structural types rather than representing a family or families of proteins. In terms of function, the proteins were selected having varying roles in transcriptional regulation. SMM slides were exposed in triplicate to each of 100 purified proteins in independent experiments and to a common epitope tag as a control. For the present study, we scored as “hits” for each protein those compounds whose average deviation from control-spot intensities exceeded a fixed threshold for statistical significance, after correction for multiple hypothesis testing (see Materials and Methods). We also eliminated compounds that bound the antibody to the control epitope tag, because these events likely do not correspond to specific binding to the 100 panel proteins. These experiments resulted in a matrix of binary hit calls for 15,252 compounds versus 100 proteins, of which 3,433 (22.5%) compounds bound at least one protein (Fig. 4). We note that this high fraction of compounds binding any protein is not unexpected with 100 parallel protein-binding assays; for example, if hit rates were 0.5% for each protein, and hit compounds were selected randomly for each of 100 proteins, we would expect $1 - (0.995)^{100} \approx 39\%$ of compounds to bind at least one protein.

We characterized the resulting 100 protein-binding profiles for each of CC, DC, and NP subsets using two measures of performance that inform how useful a small-molecule collection might be to the screening community: (i) a measure of the rate at which hits were identified from each subset, and (ii) a measure of specificity of the discovered hits from each subset, based on the number of proteins bound by a given hit. First, we examined the propensity of compounds from different sources to score positive in any protein-binding assay. A smaller proportion of NP compounds (13%; 324) were called a hit in any assay than were CC (23%; 1,415) or DC (26%; 1,694) compounds. To determine whether compounds from different sources afford different hit rates in protein-binding assays, we calculated 100 separate hit rates (one per protein) for compounds from each source. Analysis of these hit-rate distributions (Fig. 5) revealed that median hit rates are highest, and distribution least disperse, for DC. Median hit rates were intermediate for CC but with more dispersion in
their values. Median hit rates were lowest for NP, with only 7 of 100 rates for NP exceeding those in the lowest quartile for DC. Thus, compounds from different sources exhibit different group-wide behavior in protein-binding experiments, using the proteins included in this study.

This first analysis considered any binding event in establishing hit rates. To compare the binding specificity for compounds from each source, we determined the relative proportions of both highly promiscuous and highly specific compounds. To evaluate highly promiscuous compounds (Fig. 6A), we determined the relative proportion of hit compounds binding 1–5 proteins (specific), 6–24 proteins (promiscuous), or 25+ proteins (highly promiscuous). Both CC and DC harbor significantly enriched proportions of compounds that were either promiscuous (CC; 16%) or highly promiscuous (DC; 3.1%) in binding profiles. Because the synthetic pathways for DC were well-defined, we were able to categorize most (74.7%) of the 1,160 hit compounds as having come from one of 21 distinct synthetic chemistry pathways each having at least 20 members in DC (the remainder were either individual submissions or only identified by the contributing laboratory, rather than by a specific synthesis pathway).

To determine whether binding promiscuity in DC could be traced to a relatively small number of similar compounds, we tested whether the distributions of numbers of bound proteins for each of these 21 pathways could be distinguished from DC at large. We identified a single pathway, based on a spiroxindole skeleton (16), that was dramatically enriched ($p < 1.6 \times 10^{-12}$ using a Kolmogorov–Smirnov test (17) between its members and the rest of DC) in the number of proteins to which its members bound (mean = 11.4 proteins vs. mean = 2.4 proteins for the rest of DC). This insight allowed us to reanalyze promiscuity using a DC subset, termed DC’, lacking the 660 compounds derived from this synthetic pathway (of which 288 were hit compounds, the right-most 288 columns in Fig. 4). Using DC’ rather than DC in the overall comparison (Fig. 6B), we observed that only CC remained significantly enriched in promiscuous compounds and contained fewer specific members (82%) than expected by chance (i.e., if promiscuous compounds were spread across CC, DC’, and NP in proportion to the size of each group), given the lower total number of compounds. In contrast, DC (6.3%), DC’ (5.3%), and NP (2.5%) all contained proportions of promiscuous compounds lower than expected by chance (again relative to proportional representation). All other observations could be explained by statistical expectation, based on overall proportions across all three sources. This ability to identify and to eliminate problematic outliers derived from a specific synthetic pathway in DC should be useful in efforts to create optimally diverse and specific small-molecule screening collections.

To rule out whether cryptic groups of related structures might similarly explain promiscuity within CC (analogous to the spiroindoles removed from DC), we compared Tanimoto fingerprint similarities (unfolded ECFP_4) (18, 19) among the most promiscuous members of CC to those of the entire CC group and found them to be very similar (0.109 vs. 0.112 average similarity, indicating slightly more self-similarity among CC as a whole than among the most promiscuous members). We also checked the 251 most promiscuous members of CC for substructures common to at least 10% of the members and found nothing larger than an N-benzylethylamine substructure (27 instances, or 10.8%, compared with 11.7% occurrence in all of CC) in common among these compounds. These tests suggest that, in contrast to DC, the tendency to promiscuity among CC members is distributed over a large number of “chemotypes.”

We also examined the distribution of hit specificities in the CC, DC’, and NP subsets (Fig. 7). We determined the relative proportion of hit compounds binding exactly one protein (highly specific), 2–5 proteins (partially specific), or 6+ proteins (promiscuous). These groupings parallel those in our promiscuity analy-

![Fig. 3. Shape complexity of compounds from three sources. Complexity is expressed as the proportion of $sp^2$- or $sp^3$-hybridized carbon atoms that are $sp^3$-hybridized ($C_{sp3}/(C_{sp2} + C_{sp3})$); (A) whole molecules, (B) “scaffold” atoms in the molecular framework only.](image-url)

![Fig. 4. Binary hit calls for compounds from three sources against 100 proteins. Heatmap depicts presence (white) or absence (black) of a hit call for all compounds scoring as hits against at least one protein. Colored bars indicate source of compounds: CC (red), NP (green), DC (blue).](image-url)
sis, merging the two promiscuous groups into a single category and splitting the specific group into two categories. We found that CC harbors significantly enriched proportions of partially specific (28%) and promiscuous (18%) compounds and is depleted in highly specific (54%) compounds. In contrast, both DC (73%) and NP (83%) are significantly enriched in highly specific compounds, with a corresponding depletion (DC*: 6.3%; NP: 2.5%) of promiscuous compounds. Further, NP is depleted even of partially specific compounds (15%). Across all specificity categories, only the proportion of partially specific DC members was consistent with overall proportions across all three sources.

Finally, we determined whether binding trends could be connected directly with computed properties, irrespective of compound source. We categorized all 15,252 compounds into three categories of stereochemical complexity (using $C_{\text{stereogenic}}/C_{\text{total}}$), restricting our analysis to carbons in the molecular scaffold (cf., Fig. 2B): stereochemically simple ($C_{\text{stereogenic}}/C_{\text{total}} = 0$), intermediate ($0 < C_{\text{stereogenic}}/C_{\text{total}} \leq 0.25$), and complex ($C_{\text{stereogenic}}/C_{\text{total}} > 0.25$) molecules (Fig. 8). We observed that stereochemically simple molecules are significantly enriched in

![Graph](image)

**Fig. 5.** Hit-rate analysis of compounds from three sources in 100 protein-binding assays. Box-whisker plots depict second and third quartiles (blue boxes) above and below median values (red lines) with adjacent values indicating maximum nonoutlier values (black whiskers) and outlier values (red crosses); see legend at right. These data show the greatest hit rates for DC and lowest for NP (see text).

![Graph](image)

**Fig. 6.** Analysis of binding promiscuity among hit compounds. Three promiscuity categories were evaluated for compounds scored as hits against at least one protein in 100 protein-binding assays: binding to 1–5 proteins (white), 6–24 proteins (gray), 25+ proteins (black); numbers of compounds with significant enrichment (***) or depletion (*), relative to the overall proportion (far left), are indicated. These data show (A) CC members are most likely to bind 6+ proteins, NP members least likely, and DC members intermediate; (B) reevaluation after removing spiroxindole-based compounds derived from one synthetic pathway in DC to create DC (see text) suggests that much promiscuous binding among DC can be attributed to a single class of compound.

![Graph](image)

**Fig. 7.** Analysis of binding specificity among hit compounds. Three specificity categories were evaluated for compounds scored as hits against at least one protein in 100 protein-binding assays: binding to exactly one protein (white), 2–5 proteins (gray), 6+ proteins (black); numbers of compounds indicating significant enrichment (***) or depletion (*), relative to the overall proportion (far left), are indicated. These data show both DC and NP members are most likely to bind exactly one protein, while significant fractions of CC members bind at least two proteins (see text).

![Graph](image)

**Fig. 8.** Connection between binding specificity and stereochemical complexity. Four specificity categories (including non-hits) were evaluated: binding to 0 proteins (light gray), exactly 1 protein (white), 2–5 proteins (gray), 6+ proteins (black); numbers of compounds indicating significant enrichment (***) or depletion (*) relative to proportional representation are indicated. These data show stereochemically simple compounds most likely bind multiple proteins, intermediate complexity compounds most likely bind exactly 1 protein, and the most complex compounds most likely bind 0 proteins (see text).
members of this community. On the other hand, these features are notably deficient in members of compound collections currently widely used in probe- and drug-discovery efforts.

The choice of our simple stereochemical complexity descriptor for correlation with binding data reflects the motivation of the current study to link simple size-independent metrics and performance. In future studies, we will aim to test more systematically whether other descriptors, features, or physicochemical properties might be quantitatively related to such performance metrics, e.g., as judged by multivariate regression analysis. Future studies might also be more deliberate about choosing either structurally representative or diverse subsets from larger compound collections. Similarly, larger samples of compounds, different proteins, and different assay formats, especially cell-based ones, will be required to establish the generality of our conclusions.

The ability of small molecules to bind proteins selectively is just one factor that determines their value as leads in probe- and drug-discovery efforts. Not examined in this study are their ability to modulate protein functions in cells and organisms and their ease of optimization using organic synthesis. Nevertheless, these results already cast some doubt on the wisdom of the common heavy reliance on sp²-rich commercial compounds in small-molecule screening efforts. In preliminary analyses to be and recently reported elsewhere, additional differences have been identified in cellular assay contexts on the one hand and relative to natural selection (27) on the other. In the future, measuring the facility of optimizing compounds having different origins will be useful to guide even more valuable screening collections. The studies reported here constitute one step toward the goal of quantifying the biological performance of compounds from different origins and having different computed structural properties, assisting in distinguishing fact from intuition.

Materials and Methods

Compound Collections. Small molecules in CC were obtained from ChemDiv (www.chemdiv.com), Maybridge (www.maybridge.com), and TimTec (www.timtec.com). Small molecules in DC were obtained from several academic laboratories, including member laboratories of the NIGMS Centers of Excellence in Chemical Methodology and Library Development (www.nigms.nih.gov/Initiatives/CMLDL). Small molecules in NP were obtained from Analyticon (www.ac-discovery.com). In all cases, compounds were maintained by the Broad Institute compound management group prior to their use in the experiments described. Dataset S1 contains a table ("compounds") listing the simplified molecular input line entry specification (SMILES) (28, 29) representation of each compound used in this study; a table ("descriptors") indicating the category (CC, DC, or NP), structure descriptors, and ChemBank (30) names for each compound; a table ("proteins") listing the proteins used in this study; and a table ("CpdsWithProteins") of binary hit calls for all measurements. ChemBank names identify commercial sources and can be used to obtain additional information about the compounds from either ChemBank (http://chembank.broadinstitute.org) or PubChem (http://pubchem.ncbi.nlm.nih.gov). Dataset S2 contains an alternative repre-

Small-Molecule Microarrays. Each compound in the study was printed on glass microslide arrays according to published SMM protocols (13). In silico functional group analysis of compounds study revealed that 76.8% of the compounds contain at least one group that, on a functional group previously shown to undergo covalent attachment to SMMs (20) using an isocyanate surface-attachment chemistry, such as alcohols, phenols, acids, thiols, and amines. The remaining compounds may print covalently through groups not previously tested or may simply adsorb noncovalently. Proteins spanning 145 InterPro domain classifications (www.ebi.ac.uk/interpro) were obtained commercially as His6 epitope-tagged reagents for rapid identification using fluorophore-conjugated anti-His6 antibodies. Importantly, many of these proteins were exclusively tested in appropriate functional assays to ensure that they were properly folded, and all SMM experiments were performed under conditions that were developed to be compatible with protein function. Complete details of these binding assays, including SMM printing, protein-binding, wash conditions, image analysis, scoring methods, confirmatory experiments, and functional studies of hit compounds, will be described elsewhere. Dataset S1 contains a table listing the complete set of protein names used in this study, each as approved names or aliases recognized by the HUGO Gene Nomenclature Committee (http://www.genenames.org/index.html).

Data Analysis. SMM slides were scanned by a GenePix (scanner (Devices), and each SMM spot intensity was scored by its deviation from a population of vehicle-control spots on the same slide. Three replicate measurements were combined and averaged to reduce variations, normalized by the variance of corresponding vehicle-control distributions and measurement uncertainties. We called a positive hit any compound whose normalized score for protein binding exceeded the expected score for the most extreme acceptable vehicle-control outlier at a fixed statistical significance (familywise error rate p < 0.05 by Holm–Bonferroni method) (31), indicating a greater likelihood that the compound was a member of a putative hit distribution rather than the vehicle-control distribution. To ensure that hit calls were not influenced by subset-specific differences in functional groups predicted to print covalently, we compared the overall hit rates for the sets of compounds either lacking or having known-printable functional groups (20) and found these rates essentially unchanged for all three subsets (23.3% vs. 22.8% for CC, 24.7% vs. 25.7% for DC, 13.7% vs. 13.0% for NP). ECFP4 fingerprints, molecular scaffolds, molecular property counts, and derived descriptors were computed using Pipeline Pilot (Accelrys, Inc.). Statistical analyses, visualizations of distributions, and all Results figures were prepared in MATALB (The MathWorks, Inc.). Significance of enrichment or depletion of compound groups in either promiscuity or specificity groups was determined using the chi-squared test for homogeneity, with individual enrichment/depletion assignments by the method of normalized residuals (32). Dataset S1 contains a table listing the complete binary (15,252 × 100) hit calls used in the analysis.

ACKNOWLEDGMENTS. Supported by NIGMS (550-GM069721), the National Institutes of Health RoadMap (P20-HG003895), and the National Cancer Institute (N01-CO-12400). S.L.S. is an investigator at the Howard Hughes Medical Institute.


