Published in final edited form as:
J Chem Inf Model. 2011 June 27; 51(6): 1205–1215. doi:10.1021/ci1003015.

Characterizing the Diversity and Biological Relevance of the MLPCN Assay Manifold and Screening Set

Jintao Zhang[†], **Gerald H. Lushington**[‡,∥], and **Jun Huan**[*,§,∥]

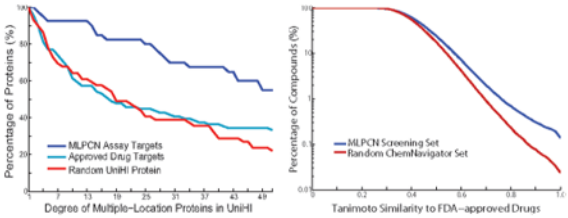
[†]Center for Bioinformatics, University of Kansas, Lawrence, Kansas 66045, United States

[‡]Molecular Graphics & Modeling Laboratory, University of Kansas, Lawrence, Kansas 66045, United States

[§]Department of Electrical Engineering and Computer Sciences, University of Kansas, Lawrence, Kansas 66045, United States

[∥]KU Specialized Chemistry Center, University of Kansas, Lawrence, Kansas 66045, United States

Abstract



The NIH Molecular Libraries Initiative (MLI), launched in 2004 with initial goals of identifying chemical probes for characterizing gene function and druggability, has produced PubChem, a chemical genomics knowledgebase for fostering translation of basic research into new therapeutic strategies. This paper assesses progress toward these goals by evaluating MLI target novelty and propensity for undergoing biochemically or therapeutically relevant modulations and the degree of chemical diversity and biogenic bias inherent in the MLI screening set. Our analyses suggest that while MLI target selection has not yet been fully optimized for biochemical diversity, it covers biologically interesting pathway space that complements established drug targets. We find the MLI screening set to be chemically diverse and to have greater biogenic bias than comparable collections of commercially available compounds. Biogenic enhancements such as incorporation of more metabolite-like chemotypes are suggested.

Introduction

Corporate pharmaceutical development has produced most of the therapeutics that are currently available to us today; but commercial development alone may not be the best vehicle for therapeutic breakthroughs because of three self-limiting factors: a) potential drug targets that are important to human health may be deprioritized if they are not financially

[*]Corresponding Author: Phone: (785)864-4620. jhuan@ittc.ku.edu.

Supporting Information. Additional experiments and analyses cited in the main text. This material is available free of charge via the Internet at http://pubs.acs.org.

Notes: The authors declare no competing financial interests.

lucrative, b) a research focus on validated drug targets may overlook genes whose selective modulation could significantly enhance knowledge of pathway dynamics relevant to therapeutic side effects and off-label applications, and c) chemical synthesis solely in the interests of drug development is often either too narrowly focused (e.g., specialized lead optimization) or too diffuse (i.e., diversity oriented synthesis driven by combinatorial convenience rather than biological relevance) to provide general chemical biology tools. Thus, a decade's worth of investment in pharmaceutical research has not produced commensurate dividends: the average annual number of novel drugs entering global markets (~26) and novel drug target discoveries (6–7) remain roughly constant[1–3], while the success rate for drug candidates seeking Food and Drug Administration (FDA)-approval has significantly decreased.[4] In response, the Molecular Libraries Initiative (MLI) was launched in 2004 to foster development chemical probes to enhance chemical biology understanding of therapeutically interesting genes and pathways and to expand the availability of small-molecule bioactivity data,[5] as are now made freely available to public and private sector researchers via the PubChem web portal (http://pubchem.ncbi.nlm.nih.gov/).[6]

The MLI aims to compensate for deficiencies in chemical biology research and genetic pathways understanding via a diverse range of large-scale bioassays (including “undruggable” targets, inconvenient for small-molecule modulation and thus underrepresented in private sector research) aimed at producing molecular probes (i.e., chemicals inducing selective modulation of phenotypically interesting biochemical or cellular functions) and providing potential insights for future therapeutic exploration.[5] While the precise metrics to qualify target druggability or undruggability are debatable, a successful target-based assay requires a functional binding site (intended for small molecule, protein, nucleotide, or other biomolecular interactions) suitable for functionally relevant small molecule complexation. Such binding sites often exhibit partial conservation across a class of homologous proteins, thus protein sequence motif recognition tools can often elucidate novel assay (and potentially druggable) candidates from understanding of known drug targets. The MLI pilot (MLSCN: the Molecular Libraries Screening Centers Network) and production (MLPCN: Molecular Libraries Probe Production Centers Network) phases have collectively conducted thousands of assays (both target- and cell-based) over a diverse array of several hundred thousand small molecules (a small subset of the estimated to exceed 10^{60} distinct small organic species[7]). Given the magnitude of this effort, it is worth evaluating the impact achieved on original MLI objectives of promoting intellectual growth in chemical biology and enhancing informational basis for novel therapeutic discovery. The first goal can be tangibly quantified: the MLI Web site currently (02/2011) reports having been a source of direct support for 345 publications during the period of 2005–2010, as a key resource supporting an additional 852 publications from 2005 to 2009, and reports the discovery and characterization of 150 molecular probes with well-documented phenotypic implications (02/2011). However, progress toward the second goal is more opaque. The aforementioned list of 1197 papers includes only four patent-related publications, and no indication is made whether these (or unpublished) studies is actually progressing toward or through clinical trials. This is not surprising given the time lag between basic research and product development, thus five years of MLI progress is an inadequate basis for such literal assessment metrics, but it is possible to compare MLI efforts with past pharmaceutical industry achievements (e.g., approved drugs and viable targets) to project the relevance of MLI pursuits toward real pharmacological development. This is what our work attempts to accomplish.

There are numerous criteria that one might place on an initiative aiming to foster new therapeutic discovery paradigms. Target-related benchmarks include whether assays generally target phenotypically important pathways (perhaps avoiding genes with excessively diverse interactions and collateral implications) with good prospects for

becoming eventual drug targets (i.e., share favorable attributes with known targets, and augment current biomedical capabilities). Thus we assess MLPCN assay targets relative to current drug targets and compare the assay screening set with known drugs and other biogenic compounds. Given our interest in analyzing specific relationships with phenotypically interesting pathways (for which the implications of cell-based assays maybe somewhat opaque), we focused our target analysis strictly on target-based biochemical assays, whose target proteins are referred to herein as *MLPCN targets*. As a basis for comparison, we used human protein–protein interaction (PPI: human protein pairs engaged in direct physical binding as reported in literature review or extrapolated from studies of orthologous proteins in other organisms.[8]) profiles to evaluate whether the MLPCN target selection would augment existing chemical genomics knowledge. Assessment of the assay screening set is geared toward determining the set's capacity for producing probes with clear, specific, and pharmacologically meaningful effects on targets. This is best satisfied with biologically relevant chemotypes without excessive promiscuity (i.e., produce 'hit' readings in assays through nonspecific mechanisms such as compound aggregation, irrelevant direct interactions with the assay reporter species and off-target binding) that may form the basis for SAR (structure–activity relationship) studies aimed at real therapeutic discovery (i.e., have physicochemical profiles and chemical functionalities resembling, but not directly duplicating, known drugs). Since the MLPCN screening set is being continually expanded to improve the likelihood of generating informative hits for all assays undertaken,[9] we chose to limit our analysis to a static set of compounds that had been tested in at least 21 of the 23 bioassays (i.e., 90% coverage) deposited into PubChem between 5/1/2009 – 7/22/2009 that screened more than 290,000 tested compounds. This produced a set of 279,768 compounds that we refer to herein as the *MLPCN screening set*. Since the body of known drugs is too small to fully reflect the chemical diversity that may emerge in novel future therapeutics, we evaluate the MLPCN screening set's composition relative to a third-party reference set of 279,768 compounds selected randomly from the ChemNavigator iResearch (http://www.chemnavigator.com) collection of commercially available compounds (herein as the random *ChemNavigator set*) from which many corporate and academic screening centers draw their compound collections. To evaluating propensity for exhibiting meaningful bioactivity, we used chemical similarity and diversity analysis techniques to assess biogenic bias and scaffold-oriented diversity of the MLPCN compounds, analyzing whether they comprised a diverse and effective facsimile for comprehensive collections like the ChemNavigator set or the entire PubChem compound collection. These analyses should measure the value of current and emerging MLI data as a research resource for basic chemical biology and future therapeutic discovery, help to elucidate strengths and deficiencies in MLPCN coverage of biochemical pathway and molecular diversity space, and potentially suggest strategies for modifying screening set composition and assay portfolio to better serve future studies.

Methods

BioAssay and Compound Data

We downloaded all bioassay and chemical structure data from PubChem (ftp://ftp.ncbi.nih.gov/pubchem/) on January 22nd, 2009. We extracted all relevant bioassay data and metadata via PERL scripts. As of 1/22/2009, there were 354 assays labeled as *primary* screening, 552 as *confirmatory*, and 230 as *summary* or other designation.

MLPCN Targets and Drug Targets in the Human PPI Network

We obtained the gene symbols and sequences of all drug target proteins from the DrugBank database[19,20] and extracted protein gene identifiers (GI) from PubChem. We manually converted GI numbers into official gene symbols by retrieving the HGNC

(http://www.genenames.org/) and NCBI Protein databases and used official gene for cross-referencing with proteins in the UniHI network and the DrugBank database.

Protein Subcellular Localization

We manually identified most subcellular protein localizations from the NCBI Gene database and the remainder from the Gene Ontology and HGNC databases. We classified MLPCN protein targets into six categories: *cytoplasm*, *extracellular*, *membrane*, *nucleus*, *organelle*, and *multiple* (proteins found in more than 2 subcellular locales, of which 50% were in cytoplasm/nucleus, 25% in cytoplasm/membrane, 22.7% cytoplasm/membrane/nucleus, and 2.3% extracellular/membrane).

Approved Drugs and Biogenic Bias

We downloaded chemical structures of all approved drugs from the DrugBank database. For natural-product-like compounds, we downloaded a subset of 89,425 natural product scaffold compounds from the ZINC database. A set of 1995 metabolites (derived from 2018 metabolites in the KEGG Compound database,[30] omitting xenobiotics) were obtained from Dr. Shoichet's group.[22] Several hundred chemically unusual compounds were removed in different analyses due to limitations in the specific chemotypes supported for descriptor calculations.

Network Topological Analysis

For our gene interaction analysis, a network can be defined as an undirected graph in which a node represents a protein, and an edge connects two nodes according to observed protein–protein interaction. The “degree” of a node is the number of edges connecting it to other nodes. Node degree distribution was computed by simple normalization of the network node distribution. The shortest path length (SPL) between two network nodes is the minimal number of consecutive edges required to connect them. In this paper, Cytoscape 2.6.2 and Network Analyzer plug-in [31,32] were used to compute SPL and degree distributions.

Statistical Tests

Wilcoxon rank-sum test, computed via the “ranksum” function in Matlab, was used to test whether medians of two sample vectors are equal, returning probability *P* of the positive answer at a given significance level (0.05 in all rank-sum tests performed in this paper).

Compound Tanimoto Similarity

We converted each compound into Daylight FP2 fingerprints and computed pairwise compound similarity via Tanimoto coefficients. The Tanimoto coefficient is not the only viable measure of compound similarity;[25] however, it is conveniently available in many chemical informatics software and compares well with other measures. E.g., for pairwise similarities across a sample of 50 random MLPCN compounds, we found Tanimoto coefficients to correlate very well with Pearson ($R^2 = 0.96$) and Dennis ($R^2 = 0.95$) coefficients, while the worst comparison (Baroni-Urbani/Buser; $R^2 = 0.81$) among a series of comparable measures[25] is still adequate for resolving the general trends analyzed herein. All fingerprint and Tanimoto computations used the free program package “open babel” (http://openbabel.org).

Compound Diversity Analysis

We evaluated the diversity of compound sets via the DiverseSolutions[26] program. Compounds were mapped into a Cartesian space defined by BCUT descriptors.[27] Using the MLPCN set to define a chemical space with substantial intercompound spatial dispersion, we mapped other compound sets into this space to contrast their relative distributions.

Results

As of 1/22/2009, the 1306 bioassays in PubChem included 1126 with at least one *active* compound (defined herein as a compound displaying assay response exceeding a threshold set by the assay practitioner to identify compounds with interesting target modulation capacity; all potentially discrepant compounds flagged by assay provider or PubChem are omitted from consideration). At this time, 151,930 compounds were identified as actives in at least one bioassay, yielding 555,859 bioassay-active/compound pairs across all assays. On average, compounds were active in 3.7 assays, and assays had 493.7 active compounds. The distributions of the bioassay (compound) count *vs.* the numbers of their active compounds (bioassays) have been investigated in Han et al.[10] In addition, 680 bioassays were considered target-based (i.e., had at least one identified target protein), and the remaining 626 were assumed to be cell-based. Figure 1a summarizes the therapeutic focus of all cell-based bioassays, finding that cancer (32%, anticancer and tumor growth inhibition), cell death (17%), and stem cell (10%) comprised the most common phenotypes. For target-based bioassays, we have found 289 distinct protein GI numbers, of which 215 had officially associated gene symbols. The distribution of target protein subcellular locations for these biochemical bioassays is shown in Figure 1b, demonstrating that the percentage of MLPCN screens focusing on membrane targets (19%) was significantly ($P < 0.001$, Chi-squared 1-degree test) smaller than among current drug targets (>50%),[11] although if one filters out target redundancy across assays or therapeutics, the difference declines somewhat (34% of MLCPN *vs.* 47% of drug targets) and becomes marginally significant ($P \approx 0.10$). To explore the molecular function of MLPCN targets, we mapped target GI numbers into EC numbers, resulting in identification of 113 distinct enzymes among 253 bioassays. The first two digits of the EC numbers enable enzymatic classification (Figure 1c), revealing hydrolases (46%) and transferases (38%) as the dominant classes of MLPCN target enzymes. Finally, the source-organism distribution of biochemical bioassays (Figure 1d) reveals that 81% of MLPCN target proteins were from human beings, while the remaining targets were from animals (6%), bacteria (6%), viruses (4%), and other miscellaneous organisms (2%).

MLPCN Targets and Approved Drug Targets

For each MLPCN target protein, we calculated global sequence similarity relative to each approved drug target via Needleman-Wunsch[12] (gap opening and extension penalties of 11 and 1, respectively) and identified the nearest neighboring drug target (i.e., highest sequence similarity). Similarly, we arbitrarily selected 500 human proteins from GenBank and identified their nearest neighboring drug targets. Distributions of the percentage of the MLPCN targets and random human proteins at each similarity score (Figure 2a) showed that 46.0% of MLPCN targets (*vs.* only 10.6% random proteins) have at least one drug-target homologue (>30% sequence identity), while 28.8% and 4.6%, respectively, were at least 90% identical to drug targets. This suggests a bias among MPLCN targets toward homology with approved drug targets, perhaps to enhance prospects probing biochemically interesting pathways. A cumulative index (relative to deposition number for target-based bioassays entering PubChem) of the percentage of MLPCN targets homologous drug targets (Figure 2b) suggests that the cumulative fraction of drug-target-like MLPCN targets fluctuated in the range of 35–60% for the first 240 bioassays and appears to be converging at about 43%. Note that to measure the effective effort that the MLPCN has committed to targets of a specific nature, we considered the full manifold of protein-based assays in this analysis without eliminating assays according to target redundancy relative to prior targets.

While 38 MLPCN targets were found to be identical to known drug targets (i.e., 100% sequence identity), 41 MLPCN targets (mostly nucleus-bound proteins) were highly distinct from them (sequence identity <15% relative to all drug targets). New insight (and potentially even new therapeutic candidates) derived from screening the 38 known drug targets is

possible but requires using a screening set containing novel compounds. The 41 distinct MLPCN targets (i.e., nonhomologous with known drug targets) afford more chance of illuminating new chemical biology, but their value is somewhat contingent on relevance to pharmacologically interesting pathways. Fortunately, 18 of them have been suggested as prospective novel drug targets in related literature (see Supplementary Table 1), and 11 of these 18 targets are cytoplasm or nucleus proteins (only four are membrane proteins), which suggests targets that are both potentially therapeutically viable and more novel.

Partitioning target proteins according to subcellular localizations (see Figure 2c) assists in identifying areas where MLPCN target selection is expanding (or just reiterating) chemical biology relative to standard pharmacological practice. Compared to established drug targets, the MLPCN target pool had significantly fewer membrane and organelle proteins but elevated ratios of nuclear and multiple-locale proteins. This apparent shift relative to cellular location distributions of known drug targets bodes well for the MLI goal of enhancing our understanding of under-represented genes.

Proteins rarely function independently but rather participate in highly interconnected protein interactome networks,[13–15] thus the differences in interaction profiles of MLPCN targets *vs.* known drug targets provides useful insight. We mapped MLPCN assay and drug targets to UniHI (unified human PPI network: >250,00 human PPIs collected from 14 major PPI sources with careful data integration and literature curation[16,17]) and identified 182 UniHI proteins for MLPCN targets and 1,035 proteins for drug targets. To gauge the potential phenotypic relevance of MLPCN targets, we calculated the shortest path lengths (SPLs) connecting each pair of MLPCN and drug targets, as compared to SPLs between two random UniHI proteins, and plotted the resulting distributions in Figure 2d. The median SPL between MLPCN and drug targets (3.393) is significantly shorter than the median UniHI SPL (4.026, $P < 10^{-262}$, Wilcoxon rank-sum test), suggesting that MLPCN targets tend to sample pathways with established therapeutically interest. Neglecting the aforementioned 38 MLPCN targets that are identical to known drug targets, we determined that 122 of the remaining 154 MLPCN targets interact directly (SPL=1) with one or more drug target(s), 29 have SPLs of 2, and only three have SPLs of 3. However, among the 122 MLPCN targets with SPL=1, most (i.e., 63.1%) have no significant homology (i.e., sequence identity <30%) with any actual drug target, while 69.0% of the 29 SPL=2 genes are similarly novel. Therefore, while most MLPCN targets probe pathways of established therapeutic interest, the focus appears to be on therapeutically unaddressed genes within those pathways. This strategy of choosing novel targets within pathways of well-established phenotypic relevance may prove to be an efficient means for identifying new alternatives to current therapeutic approaches.

MLPCN Targets and Network Degrees

From analysis that is provided in greater detail in the Supporting Information (Supplementary Figure 1a) we discovered that MLPCN targets are consistently slanted toward higher PPI degrees than drug targets and random proteins. The median degree of MLPCN targets (26.5) was found to be significantly higher than that of drug targets (12.0, $P < 10^{-5}$, Wilcoxon rank-sum test) and random UniHI proteins (9.0, $P < 10^{-8}$, Wilcoxon rank-sum test), and this trend holds even if we restrict the analysis to only high-confidence and/or medium-confidence PPIs, and if we refer to interactions reported in the CCSB-HI1 database[18] as an alternate human PPI reference. To investigate the origin of this elevated degree, we plotted the degree distributions for proteins in each subcellular location in Figure 3. For each degree k in the range of 1–51, the fraction of MLPCN targets with degree $\geq k$ was greater than that of drug targets in the cases of membrane (Figure 3a), nucleus (Figure 3b), and multiple locations (Figure 3c). Degree distributions of MLPCN targets were similar to that of drug targets in cytoplasm (Figure 3d), while extracellular and organelle MLPCN

targets generally had lower interaction degrees than the corresponding bodies of drug targets (Supplementary Figure 2a-2b). In total, the higher median degree of MLPCN targets arises primarily from cell membrane and nucleus targets. Since high degree proteins are more likely to participate in multiple pathways, their modulation is often yields biochemical implications of scientific interest. The high degree also identifies MLPCN targets as being somewhat distinct relative to the current body of drug targets, perhaps affording novel avenues for eventual therapeutics development. Conversely, there is the concern that high-degree targets may have overly complex pathway implications that are difficult to deconvolute via chemical biology modulation. However, a literature survey (see Supplementary Table 2) for the 29 MLPCN targets with UniHI degree ≥100 identifies that five of nine membrane or nucleus proteins and 13 of 20 multiple-location proteins are either existing approved drug targets or have been cited as promising targets.

Drug Likeness and Biogenic Bias of MLPCN Screening Set

We compared the MLPCN screening set, random ChemNavigator subset and 1,339 approved small-molecule drugs in DrugBank[19,20] (i.e., a collection that includes bioactive small molecule components of known drugs, but excludes biological and other larger chemical entities) from a variety of perspectives. Relative distributions of Lipinski's Rule of Five[21] parameters suggested key differences. On average, ChemNavigator compounds have larger molecular weight (Figure 4a) than MLPCN compounds and known drugs (although the latter distribution is slanted by the inclusion of smaller components in complex formulations). The distributions of octane-water partition coefficients (Figure 4b), hydrogen-bond acceptors (Figure 4c), and hydrogen-bond donors (Figure 4d) for MLPCN compounds more closely adhered to the 0–5, 0–10, and 0–5 ranges, respectively, than those of approved drugs. This obedience to Rule of Five characteristics is somewhat counterintuitive since the MLPCN has no direct mandate to focus on orally available compounds but may be a consequence of having procured a substantial majority of the collection from commercial vendors accustomed delivering Lipinski-compliant species. Loosening the Rule of Five tendency in in new acquisition MLPCN compound acquisitions may provide a justifiable mechanism for achieving greater diversity among potentially bioactive chemotypes.

Biogenic bias is the predisposition toward substances (e.g., metabolites, natural products, and peptides) that are produced *in vivo* and is considered a “good” bias in screening set design.[22] Metabolites are small molecule components of primary metabolism, comprising many scaffolds similar to existing drugs.[23] Even unconventional metabolites such as glycans[33] and lipids[28] are accruing attention as core scaffolds from which to devise prospective therapeutics. Natural products have been optimized via natural selection for optimal interactions with biological macromolecules.[24] Metabolites and natural products both provide excellent building blocks in novel bioactive molecule design, while peptides are often a source of inspiration for the formulation of peptide-like species. It is thus useful to assess how biased MLPCN sets are toward metabolites, natural products, and peptide. To do so, we characterized each MLPCN and ChemNavigator compound by identifying its nearest neighbors (i.e., the compounds with highest Tanimoto similarity) in a set of approved drugs, metabolites, natural-product-like, and peptide-like compounds and plotted the distributions as percentages of compounds at varying similarity scores (see Figure 5a-d). In all cases, MLPCN compounds were more similar to approved drugs and natural-product-like compounds than were random ChemNavigator compounds. Five-fold more MLPCN compounds than ChemNavigator compounds had a mean Tanimoto similarity of $T_c = 0.83$ (0.94) to natural-product-like compounds (approved drugs). MLPCN compounds were also more similar to metabolites at all T_c values but by a smaller margin. Interestingly, fewer peptide-like compounds are found in MLPCN screening set than random ChemNavigator set. From these distributions one may surmise that the addition of more scaffolds exhibiting

common metabolite substructures or peptide-like constructs into the MLPCN screening set may further augment coverage of likely bioactivity chemotypes.

Compound Diversity of MLPCN Screening Set

Key steps in designing an optimal screening library entail assessing chemical space coverage, structural novelty, pharmaceutical, and biological relevance compared to important reference compound sets such as approved drugs, metabolites, natural product analogs, and commercially available compounds. To compare MLPCN screening set with ChemNavigator collection diversity, we modeled chemical space coverage via the Tripos DiverseSolutions[26] program by mapping each compound from both sets into an N-dimensional space defined by BCUT molecular descriptors.[27] We used the autoselect option to choose three descriptors to best define a 3D chemical space for the MLPCN screening set according to optimal compound dispersion across Cartesian space. We selected the top two descriptors to make a 2D chemical descriptor space and mapped the two sets into this space, partitioning each axis into 600 equal bins to produce the heatmap images in Figure 6a (MLPCN) and 6b (ChemNavigator). Cell-fraction distributions of the compounds over a coarser 10 × 10 grid are shown in Figure 6c. These figures suggest substantial spatial overlap of MLPCN and ChemNavigator distributions was substantial.

In addition, we also mapped approved drugs, metabolites, natural product analogs, and peptides into the same space, with distributions shown in Figure 7a-d, respectively. The distribution of approved drugs, although sparse, was heavily concentrated in areas with a reasonable MLPCN compound density, suggesting a predisposition of MLPCN compound providers for emulating approved drugs. Natural product scaffold distribution also mirrored approved drugs but had a significant fraction of scaffolds in regions underrepresented by MLPCN compounds. A substantial fraction of peptides and metabolites occupied regions where few or no MLPCN compounds can be found (and vice versa), suggesting that metabolite-like and peptide-like scaffolds are underrepresented in the MLPCN screening set and could comprise a biochemically useful augmentation.

Discussion

Beyond the basic objective of producing chemical probes for studying the functions of genes, cells, and biochemical pathways,[9] an original mandate of the MLPCN program was to provide a knowledgebase to support drug discovery toward important biomedical objectives.[5] Our study aims to probe the utility of MLPCN screening data toward these ends by addressing a) the relationships between MLPCN targets and existing drug targets in human PPI networks, b) differences between the MLPCN screening set versus commercially available compounds and approved drugs, and c) diversity and biological relevance of the MLPCN screening set. Although not directly aligned with current MLPCN mandates, this analysis may prove useful in gauging MLPCN progress as a vehicle for fostering chemical biology research and new therapeutics discovery.

Our sequence analysis reveals that over 40% of MLPCN targets are significantly homologous with at least one established drug target, with approximately the equal numbers of MLPCN targets being identical to known drug targets as those being completely nonhomologous. This appears to strike a plausible balance between *de novo* targeting and possible discovery of novel modulators for established targets. From protein interaction network analysis, we found that MLPCN targets (especially membrane, nucleus, and multiple-location proteins) had a median interaction degree significantly greater than both the UniHI median and that of approved drug targets. This is a key metric by which MLPCN targets differ from the body of established drug targets, but it is unclear whether the high pathway interactivity degree of MLPCN targets makes them more or less scientifically

interesting; in one sense interactive complexity is a measure of likely phenotypic relevance but may also hinder deconvolution of the physiological implications that might arise from therapeutic modulation. Interestingly, a literature survey confirmed that many of the highest-degree MLPCN targets correspond to approved drug targets or identified therapeutic prospects. This suggests that phenotypic advantages of pathway complexity may outweigh practical challenges toward prospective therapeutic applications.

Our network analysis also revealed that MLPCN targets are much more likely to have direct physical interactions with established drug targets than is the case for randomly selected genes within the UniHI human PPI database. Interestingly, these closely associating genes are largely nonhomologous with known drug targets, thus MLPCN target space is populated with novel, hitherto untargeted species participating in therapeutically relevant pathways. This bodes well for promoting novel therapeutic discovery.

Our chemoinformatic analyses suggest substantial biological relevance inherent in the MLPCN screening set, providing a solid basis for basic research and future drug discovery, but there may be key opportunities for further enhancing the biogenic nature of the set. MLPCN compounds follow the Rule of Five more closely than approved drugs; discarding this apparent Lipinski bias could permit greater flexibility in augmenting the set with more biologically relevant compounds. Chemical similarity profiling of the MLPCN screening set versus a random subset of ChemNavigator compounds suggested that the former has more representatives that are closely related to approved drugs and scaffolds based on natural products, while representation of metabolite-based scaffolds was similar in the two sets, and the MLPCN set had fewer peptide-like compounds. By scanning for regions of chemical space occupied by scaffolds based on metabolites[23] and natural products,[24] and peptides, we evaluated the biogenic bias of the MLPCN screening set and determined that all three biogenic classes have distribution into regions where MLPCN compounds are absent but approved drugs are represented. Augmenting the MLPCN set with representatives from these underrepresented biogenic scaffolds could significantly enhance biologically relevant chemical-space coverage. For example, the value of screening sets with greater representation of metabolite scaffolds has been demonstrated by Dobson et al.[23] who determined that the manifold of known FDA-approved therapeutics bears significantly greater structural and physicochemical similarity to the body of known metabolites than to the average Lipinski-compliant synthetic organic compounds found in many screening sets, although species excessively similar to known intermediary metabolites are often metabolically unstable and thus impractical for therapeutic application.

In conclusion, this paper attempts to objectively assesses the MLPCN program as a resource for enhancing chemical biology and drug discovery, probing the relative novelty of target selection, the likelihood that these targets will prove scientifically or therapeutically interesting, the relative chemical diversity inherent in the assay screening set, and the extent of biogenic bias in the screening set has that is likely to modulate interesting biochemistry. MLPCN target selection appears to strike a reasonable balance between established targets above which more can be learned, novel targets that probe therapeutically established pathways, and highly interactive homologues to known drug targets that have not themselves yet been targeted. The MLPCN screening set is found to overlap a reasonable fraction of the chemical space occupied by available drug-like small molecules, with greater biogenic bias than a comparable-sized set of commercially available compounds; however, some areas for prospective biogenic enhancement (through which prospective impact of screens might theoretically be bolstered) are proposed for consideration should the MLPCN pursue selective procurement of novel subsets from corporate compound collections or for recruiting compound donations from academic sources.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work has been supported by the National Science Foundation under Grant No. IIS 0845951 and the KU Specialized Chemistry Center (NIH U54 HG005031). The authors thank Dr. Matthias Futschik and Dr. Brian K. Shoichet for providing us the full UniHI database file and the collection file of metabolites, respectively. The authors would like to thank Jeffrey Aubé for scientific feedback and Nora Wallace for assistance in the preparation of this manuscript.

References

1. Adams C, Brantner V. Estimating the cost of new drug development: is it really 802 million dollars? Health Aff (Millwood). 2006; 25(2):420–428. [PubMed: 16522582]
2. Yildirim MA, Goh KIl, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol. 2007; 25:1119–1126. [PubMed: 17921997]
3. Cokol M, Iossifov I, Weinreb C, Rzhetsky A. Emergent behavior of growing knowledge about molecular interactions. Nat Biotechnol. 2005; 23:1243–1247. [PubMed: 16211067]
4. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discovery. 2004; 3:711–716.
5. Austin CP, Brady LS, Insel TR, Collins FS. NIH Molecular Libraries Initiative. Science. 2004; 306:1138–1139. [PubMed: 15542455]
6. Zerhouni E. The NIH Roadmap. Science. 2003; 302:63–72. [PubMed: 14526066]
7. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16:3–50. [PubMed: 8788213]
8. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics. 2005; 21:2076–2082. [PubMed: 15657099]
9. Oprea TI, Bologa CG, Boyer S, Gurpan RF, Glen RC, Hopkins AL, Lipinski CA, Marshall GR, Martin YC, Ostopovici-Halip L, Rishton G, Ursu O, Vaz RJ, Waller C, Waldmann H, Sklar LA. A crowdsourcing evaluation of the NIH chemical probes. Nat Chem Biol. 2009; 5(7):441–447. [PubMed: 19536101]
10. Han LY, Wang YL, Bryant SH. A survey of across-target bioactivity results of small molecules in PubChem. Bioinformatics. 2009; 25(17):2251–5. [PubMed: 19549631]
11. Drews J. Drug Discovery: A Historical Perspective. Science. 2000; 287:1960–1964. [PubMed: 10720314]
12. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970; 48(3):443–453. [PubMed: 5420325]
13. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5:101–113. [PubMed: 14735121]
14. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout A, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004; 430:88–93. [PubMed: 15190252]
15. Jeong H, Mason S, Barabasi AL, Oltvai Z. Lethality and centrality in protein networks. Nature. 2001; 411:41–42. [PubMed: 11333967]
16. Chaurasia G, Iqbal Y, Hanig C, Herzel H, Wanker EE, Futschik ME. UniHI: an entry gate to the human protein interactome. Nucleic Acids Res. 2007; 35:D590–D594. [PubMed: 17158159]
17. Chaurasia G, Malhotra S, Russ J, Schnoegl S, Hanig C, Wanker EE, Futschik ME. UniHI 4: New tools for query, analysis and visualization of the human protein-protein interactome. Nucleic Acids Res. 2009; 37(Database issue):D657–D660. [PubMed: 18984619]
18. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton G, Llamosas E,

Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamma L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437:1173–1178. [PubMed: 16189514]

19. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36(Database issue):D901–906. [PubMed: 18048412]
20. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. Nucleic Acids Res. 2006; 34(Database issue):D668–672. [PubMed: 16381955]
21. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Delivery Rev. 1997; 23:3–25.
22. Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK. Quantifying biogenic bias in screening libraries. Nat Chem Biol. 2009; 5(7):479–483. [PubMed: 19483698]
23. Dobson PD, Patel Y, Kell DB. Metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. Drug Discovery Today. 2009; 14(1–2):31–40. [PubMed: 19049901]
24. Ertl P, Roggo S, Schuffenhauer A. Natural product-likeness score and its application for priorization of compound libraries. J Chem Inf Model. 2008; 48:68–74. [PubMed: 18034468]
25. Haranczyk M, Holliday J. Comparison of Similarity Coefficients for Clustering and Compound Selection. J Chem Inf Model. 2008; 48(3):498–508. [PubMed: 18293953]
26. Sybyl. Version 7.3.3. TRIPOS, Inc.; St Louis, MO 63144: 2007.
27. Pearlman RS, Smith KM. Software for chemical diversity in the context of accelerated drug discovery. DrugsFuture. 1998; 23:885–895.
28. Raulin J. Recent Developments in Lipid Drugs. Mini-Rev Med Chem. 2005; 5(5):489–498. [PubMed: 15892690]
29. Irwin JJ, Shoichet BK. ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005; 45:177–182. [PubMed: 15667143]
30. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27–30. [PubMed: 10592173]
31. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13(11):2498–2504. [PubMed: 14597658]
32. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics. 2008; 24(2):282–284. [PubMed: 18006545]
33. Kulkarni AA, Weiss AA, Iyer SS. Glycan-based high-affinity ligands for toxins and pathogen receptors. Med Res Rev. 2010; 30(2):327–393. [PubMed: 20135686]

Abbreviations List

MLPCN	Molecular Libraries Probe Production Centers Network
MLI	Molecular Libraries Initiative
MLSCN	Molecular Libraries Screening Centers Network
FDA	Food and Drug Administration

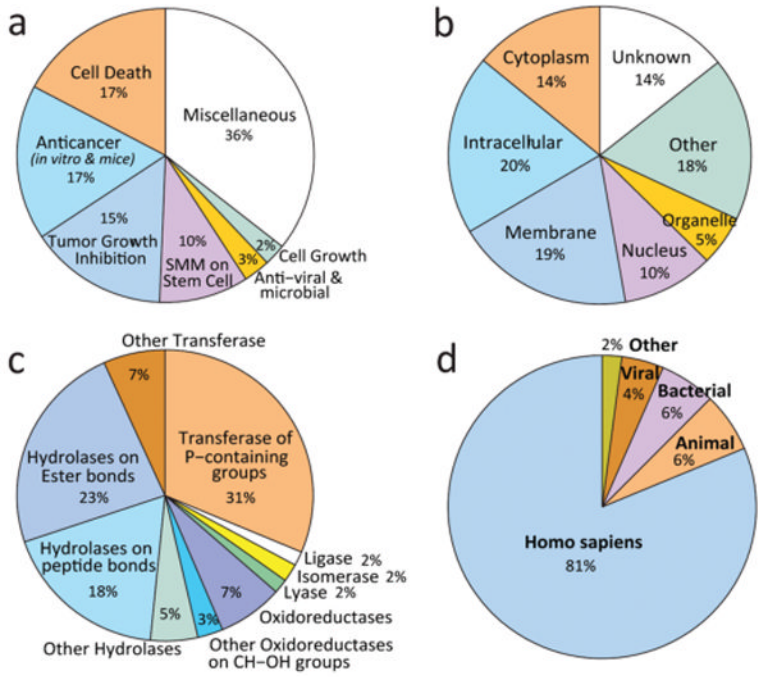


Figure 1.

BioAssay data characteristics. (a) Phenotypic distribution 626 cell-based bioassays. (b) Target subcellular localization of the 680 biochemical bioassays. (c) Functional class distribution of the 113 MLPCN target enzymes. (d) Source organism distribution for MLPCN biochemical assay targets.

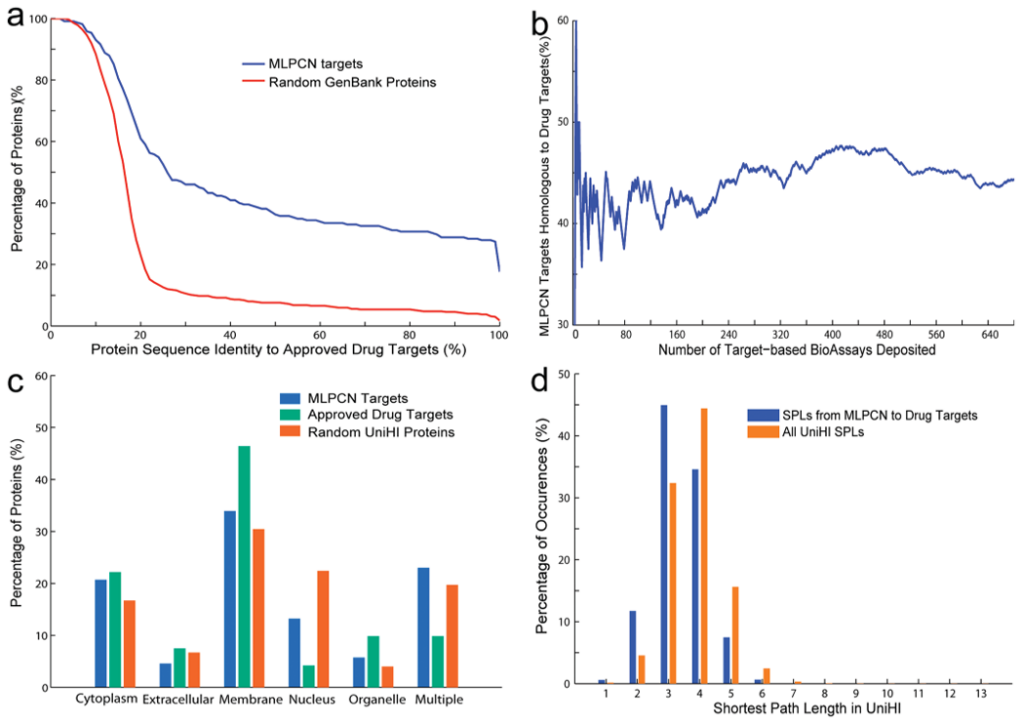


Figure 2.

Comparison of MLPCN assays vs drug targets. (a) Fraction of MLPCN targets and random human proteins as a function of sequence identity to drug targets. (b) Change in percentage of MLPCN assay targets homologous (sequence identity ≥30%) to drug targets over time. (c) Relative subcellular distribution of MLPCN targets, drug targets, and random UniHI proteins. (d) Comparison of SPLs between MLPCN and drug targets vs SPLs between all UniHI proteins.

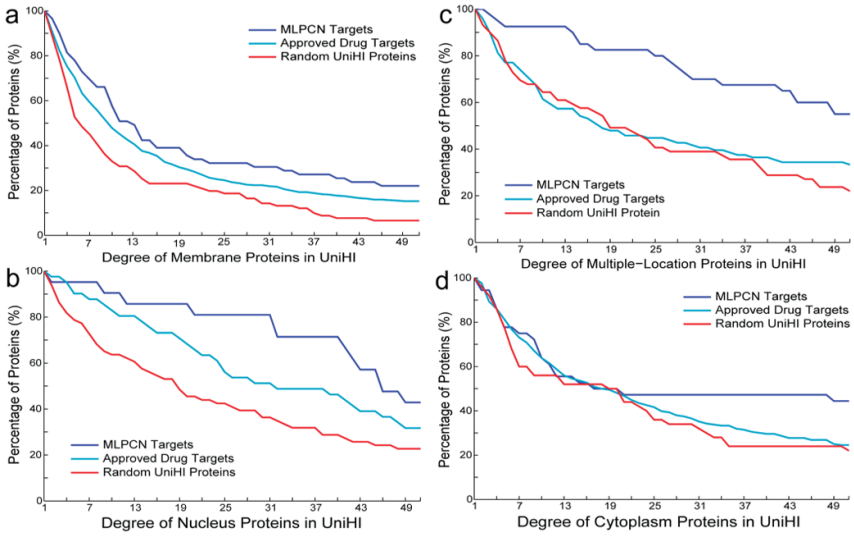


Figure 3.

Distributions of the percentages of MLPCN targets, drug targets, and random UniHI proteins at varying UniHI degrees, as classified into (a) membrane, (b) nucleus, (c) multiple-location, and (d) cytoplasm proteins. A protein is classified into category (c) if it can be located in two or more of the following subcellular locations: cytoplasm, extracellular, membrane, nucleus, and organelle.

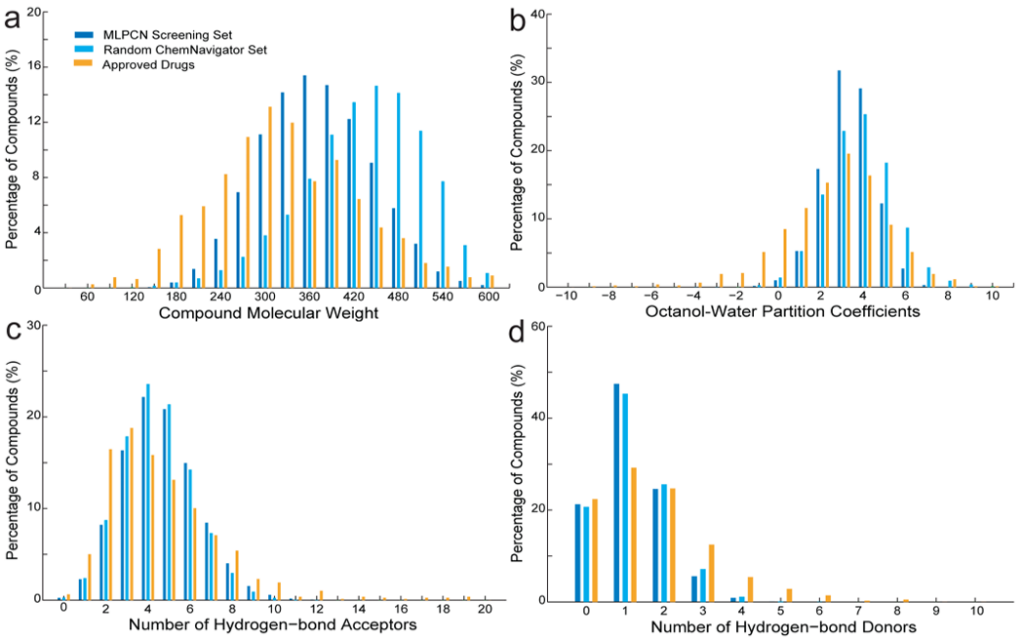


Figure 4.

Distributions of the MLPCN screening set (blue), approved drugs (yellow), and random ChemNavigator set (cyan) for (a) molecular weight, (b) octane-water partition coefficient, (c) H-bond acceptors, and (d) H-bond donors.

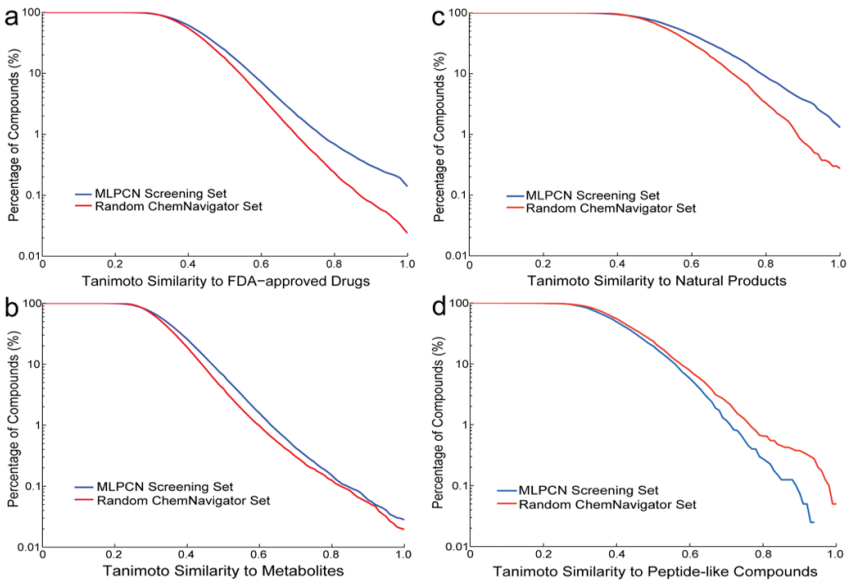


Figure 5.

Drug likeness and biogenic bias of the MLPCN screening set (blue) vs random ChemNavigator set (red) using the distributions of their nearest neighboring (a) approved drugs, (b) metabolites, (c) natural products, and (d) peptide-like compounds at each Tanimoto similarity score.

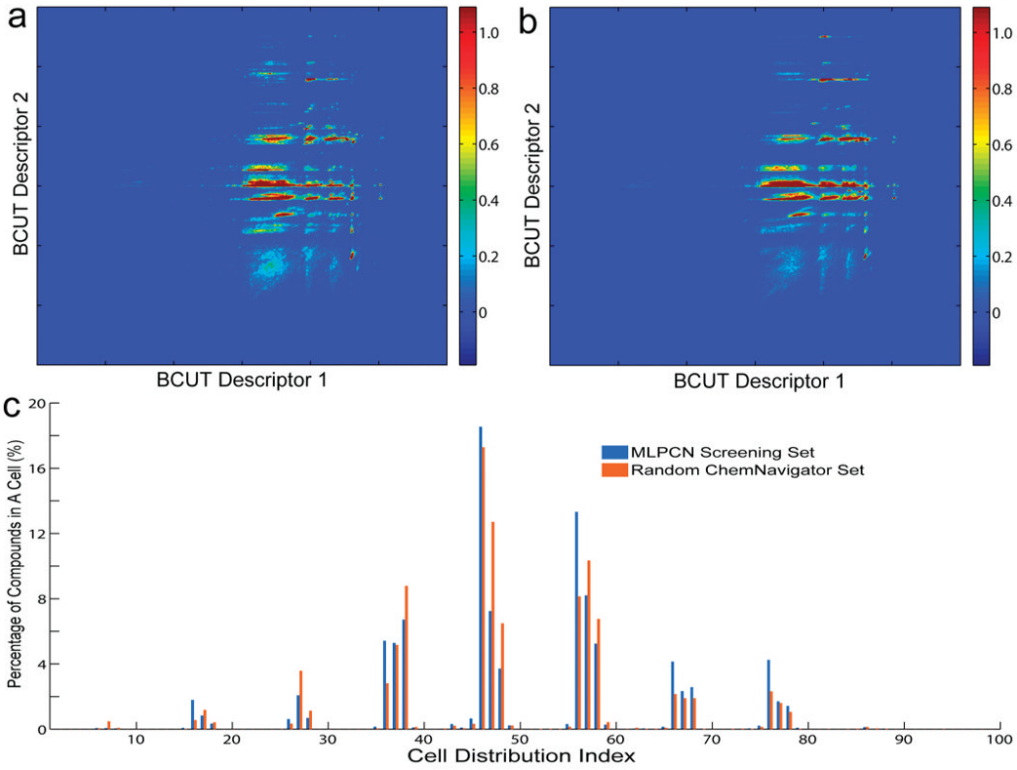


Figure 6.

Distributions of (a) the MLPCN screening set and (b) the random ChemNavigator set in a descriptor space partitioned into 600 × 600 cells, respectively. Cells are colored according to per-cell fraction via the color bar (scaled as 1.0 → 0.0001). For better visibility, images are blurred with 30% on original pixels and equal contribution of neighboring pixels. (c) Specific distributions for MLPCN and ChemNavigator compound fractions for a more coarse partitioning over 10 × 10 = 100 cells. “Cell distribution index” is calculated as 10(row number − 1) + column number. Axis labels “BCUT Descriptor 1” and “BCUT Descriptor 2” refer to “BCUT_haccept_burden_000.900_R_H” and “BCUT_tabpolar_burden_000.500_R_H”, respectively.

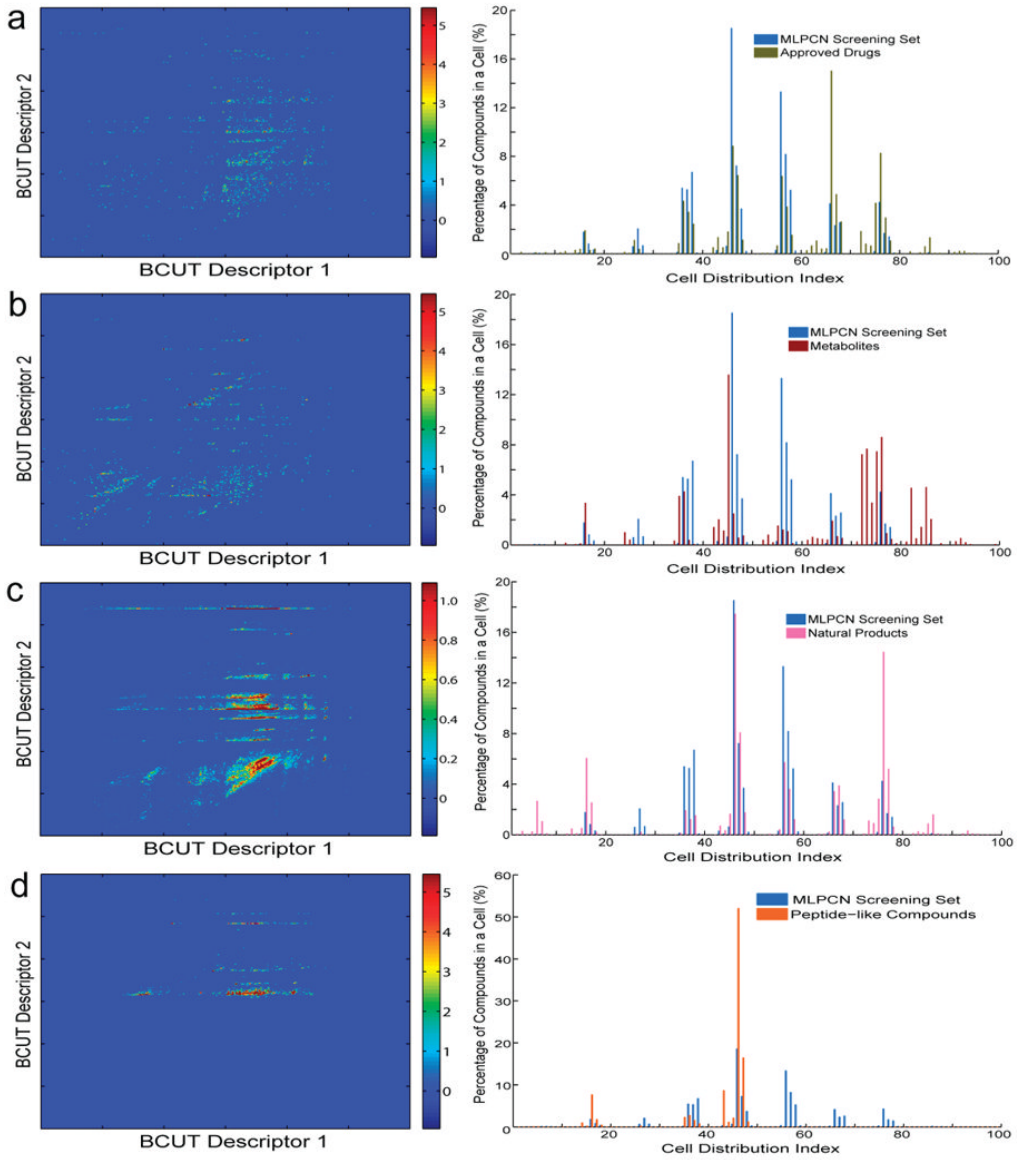


Figure 7.

Distributions of (a) approved drugs, (b) metabolites, (c) natural products, and (d) peptide-like compounds in the same MLPCN descriptor space partitioned into 600*600 cells, respectively. Cells are colored according to per-cell fraction via the color bar (scaled as 1.0 → 0.0001). For better visibility, images are blurred with 30% on original pixels and equal contribution of neighboring pixels. Bar charts on the right show the specific distributions of the fractions of these four sets of compounds in each cell using a coarse partitioning of 10 × 10 = 100 cells. Axis labels identical to Figure 6.