JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

The Purchasable Chemical Space: A Detailed Picture

Xavier Lucas, Björn A. Grüning, Stefan Bleher, and Stefan Günther*

Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Hermann-Herder-Str. 9, D-79104 Freiburg, Germany

Supporting Information

ABSTRACT: The screening of a reduced yet diverse and synthesizable region of the chemical space is a critical step in drug discovery. The ZINC database is nowadays routinely used to freely access and screen millions of commercially available compounds. We collected ~125 million compounds from chemical catalogs and the ZINC database, yielding more than 68 million unique molecules, including a large portion of described natural products (NPs) and drugs. The data set was filtered using advanced medicinal chemistry rules to remove potentially toxic, promiscuous, metabolically labile, or reactive compounds. We studied the physicochemical properties of this compilation and identified millions of NP-like, fragment-like, inhibitors of protein−protein interactions (i-PPIs) like, and drug-like compounds. The related focused libraries were subjected to a detailed scaffold diversity analysis and compared to reference NPs and marketed drugs. This study revealed thousands of diverse chemotypes with distinct representations of building block combinations among the data sets. An analysis of the stereogenic and shape complexity properties of the libraries also showed that they present well-defined levels of complexity, following the tendency: i-PPIs-like < drug-like < fragment-like < NP-like. As the collected compounds have huge interest in drug discovery and particularly virtual screening and library design, we offer a freely available collection comprising over 37 million molecules under: http://pbox.pharmaceutical-bioinformatics.org, as well as the filtering rules used to build the focused libraries described herein.

INTRODUCTION

The chemical space comprises the virtually infinite set of possible organic compounds.[1] Even though it certainly contains unknown bioactive small molecules, their identification by systematic screening is inconceivable. Current technology accounts for the evaluation of up to a few million molecules in high-throughput screening (HTS) and even more in its in silico fashion.[2−4] In either case, the screening of a preselected reduced yet diverse region of the chemical space enriched with promising compounds is a critical step in drug discovery. During the last decades, several techniques and semiempirical rules have been proposed to efficiently navigate the chemical space, including the selection of natural products (NPs) resulting from billions of years of evolution and compounds resembling marketed drugs.[5−7] Such cheminformatics tools have been adopted by the drug discovery community, including pharmaceutical companies and academia.[8,9]

Navigating the chemical space, however, is further restricted by the feasibility of obtaining putatively active small molecules. Purification of NPs from living organisms,[10] organic and enzymatic synthesis of compounds,[11−13] and combinatorial chemistry and biosynthesis pursuing novel derivatives[14,15] are time-consuming strategies. A fast and significantly cheaper alternative to overcome this issue is the mining of readily commercially available compounds, e.g. from the ZINC database.[16] This resource is nowadays routinely used as a primary source of commercial compounds, and comprises millions of molecules ready-to-dock. The versatility of ZINC has been largely proven by the discovery of many bioactive compounds acting on a broad target spectrum.[3,17−20] Nonetheless, the diversity and coverage of the ever growing purchasable space requires a detailed, critical, exhaustive assessment in order to determine its suitability and maximize its usability in drug discovery.

In this manuscript, we have collected many millions of commercially available small molecules from companies' catalogs and the ZINC database using the freely available cheminformatics platform ChemicalToolBoX,[21] filtered them using advanced medicinal chemistry protocols such as Pan Assay Interference Compounds (PAINS),[22] aiming at the exclusion of potentially reactive, metabolically labile, promiscuous, or toxic compounds, and classified the resulting data set into several focused ligand libraries, e.g. drug-like, inhibitor of protein−protein interactions (i-PPIs) like, and NP-like. We have analyzed the physicochemical properties and compound diversity of the collected chemicals, carried out a detailed scaffold analysis, compared them to reported data sets of NPs and marketed drugs, and studied the molecular complexity of the different focused libraries in terms of stereogenicity and hybridization. A growing interest in this property has emerged in the drug discovery field, mainly due to its reported correlation with biological target specificity and aqueous solubility.[23−25] Moreover, we have recently shown that drug targets recognize substrates with diverse, characteristic degrees of stereogenicity and that complex molecules are particularly

Received: December 12, 2014
Published: April 20, 2015

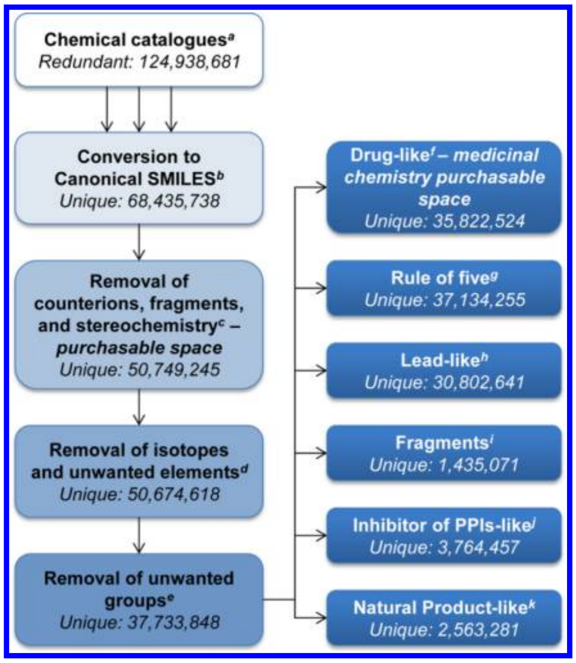
DOI: 10.1021/acs.jcim.5b00116

relevant to inhibiting low-druggability binding sites.[26] Hence, a study of stereogenic and shape complexities in library design is presented here.

METHODS

Selection of Commercially Available Compounds. The workflow shown in Scheme 1 was used to query and process the catalogs of 115 vendor companies from all over the world plus the ZINC database[16] (Table S1). The resulting data set contained ~125 million molecules, 68 million of which were unique. Chemicals not offered for direct download were stored locally and subsequently included in the workflow. Molecule conversions and comparisons throughout the manuscript are based on canonical SMILES at pH 7.0 using canonical signatures for dative bonds and nonstereogenic labeling generated using Open Babel 2.3.1,[31] as implemented in the ChemicalToolBoX.[21] The same platform was used for the prediction of physicochemical properties.

Scheme 1. Workflow Used to Collect, Filter, and Partition the Purchasable Space



[a]A complete listing of the 115 collaborating vendors is shown in Table S1. [b]Canonical SMILES were generated at pH 7.0 using canonical signatures for dative bonds. [c]Counterions, stereochemical information, and molecules with less than six heavy atoms were removed. [d]Organic molecules with isotope labeling or consisting of chemical elements different from H, C, N, O, S, F, Cl, Br, or I were removed (*grep*-based filtering rules shown in Table S2). [e]Filtering rules based on advanced medicinal chemistry rules and PAINS[22] were applied (details in Tables S3–S5). [f]Selection based on quantitative estimate of drug-likeness (QED) score[27] ≥ 0.35. [g]Molecules with no more than one violation of $MW \leq 500$, $ALOGP \leq 5$, $HBD \leq 5$, and $HBA \leq 10$.[6] [h]Molecules with $MW \leq 460$, $-4 \leq ALOGP \leq 4.2$, $HBD \leq 5$, $HBA \leq 9$, $RINGS \leq 4$, and $ROTB \leq 10$.[7] [i]Molecules with $MW < 300$, $ALOGP \leq 3$, $HBD \leq 3$, $HBA \leq 3$, $ROTB \leq 3$, and $TPSA \leq 60$.[28] [j]Molecules with $400 \leq MW \leq 700$, $1.5 \leq ALOGP \leq 6.5$, $4 \leq HBA \leq 9$, and $3 \leq RINGS \leq 6$; adapted from the work of Morelli et al.[29] [k]Molecules with natural product (NP) likeness scorer system[30] ≥ 0.0.

Preparation of the Reference Set of NPs. We generated a large collection of reference NPs by merging three manually curated NPs databases: the Traditional Chinese Medicine Database @ Taiwan,[32] the Dictionary of Natural Products,[33] and the database of streptomycetes-derived metabolites StreptomeDB.[34] The resulting reference set contained 227 241 unique NPs and is referred to as "reference NPs" throughout the manuscript.

Application of Advanced Chemical Filters to the Purchasable Space. Compound exclusion based on structural features is challenging for large data sets.[35] We took advantage of the multiprocessor capabilities of the ChemicalToolBoX[21] to include filters for compounds containing chemical alarms found as potentially reactive, metabolically labile, toxic, or repeatedly interfering in assays (PAINS).[22] To increase usability and performance, many SMARTS were converted to equivalent SMILES and processed using the UNIX command tool *grep*. A complete listing of applied filters as SMILES and SMARTS descriptors is included in Tables S2–S5. Their application to the collected purchasable space identified 13 million compounds with chemical alarms (25.5%, Scheme 1).

Scaffold Diversity Analysis. The Scaffold Decomposition tool included in Canvas 2.1 (Schrödinger, LLC, New York, NY, USA) was used to process the focused libraries of i-PPIs-like, fragment-like, and NP-like and the reference NPs and approved drugs from DrugBank[36] sets and to collect distinct level 0 and level 1 scaffolds. A small molecule can comprise one or more level 0 scaffolds, i.e. root cyclic or polycyclic structures, as well as one or multiple level 1 scaffolds, i.e. substructures containing two chemically equivalent or different level 0 entities connected by an aliphatic or functionalized linker. A cumulative scaffold frequency plot was obtained by representing the cumulative percentage of scaffolds against the sorted cumulative scaffold frequency with reference to the total number of molecules, i.e. the percentage of molecules that contain a particular scaffold.[37] A logarithmic scaffold frequency plot was also generated by plotting the scaffold frequency against the sorted scaffolds in log–log scale to determine the overall redundancy of the data set.

RESULTS AND DISCUSSION

Selection of Commercially Available Small Molecules. We collected ~125 million commercially available compounds, comprising over 68 million unique chemicals, using an automated workflow (Scheme 1 and Methods). In contrast to other large compilations, e.g. the licensed Chemical Abstracts Service (CAS)[38] and the ChemNavigator's iResearch Library,[39] the purchasable space described herein can be transparently reproduced by any user. Even though it cannot be ensured that each molecule included in this collection is ready for delivery, it gives a general overview of the millions of molecules that are offered by chemical vendors. We received the agreement from 115 vendors to access their catalogs and to perform the presented studies (Table S1). These included both large-scale and focused collections. For example, Enamine Ltd. (Kiev, Ukraine) offered the largest data set for HTS, which contained over 35 million molecules with roughly half of them being uniquely available, while Cayman Chemical Company (Ann Harbor, MI, USA) provided a small collection of a few thousand naturally derived compounds. Some suppliers, such as Life Chemicals Inc. (Ontario, Canada), had many ready-to-use predefined target-focused libraries. Novel synthetic chemistry compounds were collected from specialized companies: Squarix GmbH (Marl, Germany) offered a small, diverse, rare compilation of small molecules. Many companies additionally

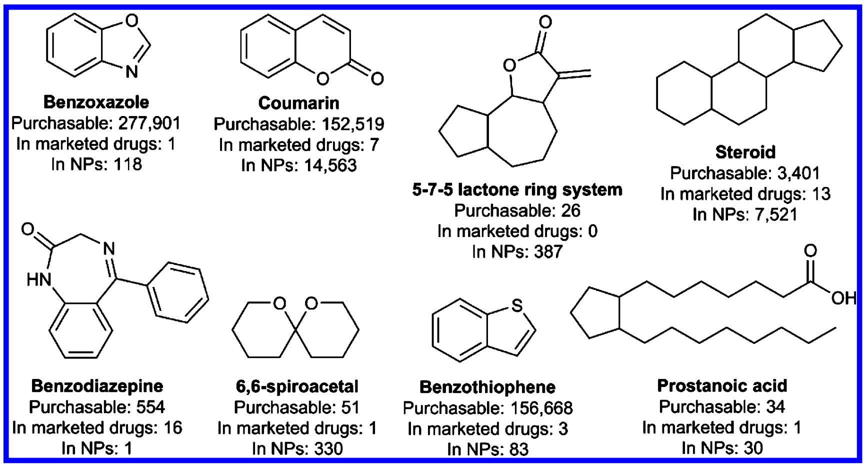


Figure 1. Examples of chemotypes found in marketed drugs and NPs defined by Welsch et al.[40] Their recurrence in bioactive molecules suggests that they are promising starting points for drug discovery.

provide custom synthesis services that can be used to further extend the reachable chemical space.

Purchasable Space As a Source of NPs and Repurposed Drugs. Traditionally, NPs have been the single most productive source of lead compounds for the development of drugs even after the inclusion of HTS techniques.[5] However, their extraction from natural sources or their organic synthesis remain major drawbacks for pursuing large-scale drug discovery campaigns. Their commercial availability would significantly ease and accelerate the process. Thus, we collected NPs contained in the purchasable space. Out of the 227 241 naturally derived molecules contained in large databases (see Methods), 81 182 compounds (36%) could be identified in commercial catalogs as exact matches.

Another technique used for the identification of bioactive molecules is drug repurposing. During the past decade, this approach has gained increasing attention as a drug discovery alternative aiming at the therapeutic switching of known bioactive molecules.[41] Cancer therapy has specifically benefited from this approach, e.g. it led to the unexpected discovery of anticancer activity in compounds such as the immunosuppressant rapamycin and the antiepileptic agent valproic acid.[42] To assess the potential of the purchasable space in drug repurposing, we selected commercially available molecules within the approved drugs subset of DrugBank.[36] Out of 1493 approved drugs, a total of 1360 (91%) were identified as commercially available.

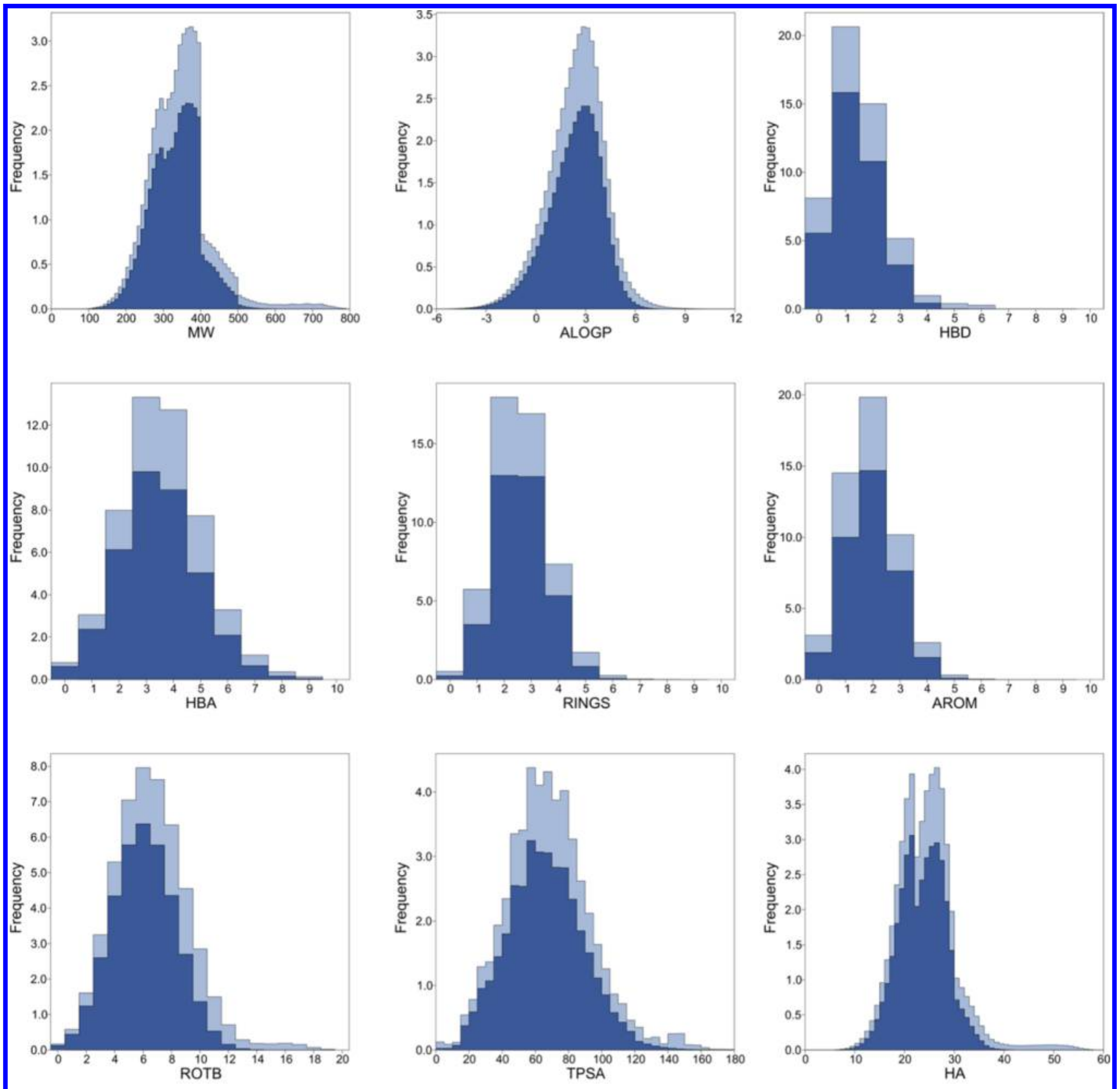
Figure 1 shows a collection of chemotypes found in marketed drugs and NPs,[40] and the number of compounds containing each scaffold in the purchasable space. Some bioactive structural motifs prevalent in NPs, such as the anti-Leishmanial and -Trypanosomal 5−7−5 lactone ring system, are scarcely found or inexistent in marketed drugs. Cumbersome chemical synthesis routes or elaborative extraction methods certainly complicate their biological characterization.[40] Their acquisition from commercial catalogs represents an efficient alternative to study many of them now.

Drug-Likeness Profiling of Commercially Available Compounds. As many approved drugs could be identified within the collected purchasable space, containing over 50 million molecules, this data set was further analyzed in terms of common physicochemical descriptors (Figure 2, sky blue). This study provides an objective, unbiased overview of the properties of commercially available molecules which, taken up to the 90th percentile, pass the ubiquitously used “rule of five” (ro5).[6] As the properties of this semiempirical rule are consistent with the physicochemical profile of oral drugs,[43] the majority of the compounds collected herein present, in a general perspective, promising drug-like properties. A more detailed inspection of the distributions also suggests that companies apply strict selection rules to cover desirable regions of the chemical space, e.g. the median molecular weight (MW) is 343 Da, with only 4% of molecules presenting a MW over 500 Da. This value contrasts with the larger 387 Da recently described by Zuegg and Cooper after analyzing their own data set of unique commercially available chemicals.[44] On the one hand, the difference could arise from analyzing a significantly smaller data set of ∼8 million molecules. On the other hand, we observed a drastic reduction of molecules with MW above 400 and 500 Da in our compilation, which derives into a skewed representation of the property. We could also observe a remarkable decrease in the count of molecules bearing exactly 22 heavy atoms (HAs). Providers included in the study indicated that these effects resulted likely from the application of diverse filters for drug-like and fragment molecules during the generation of chemical catalogs.

Application of Advanced Chemical Filters to the Purchasable Space. One of the main applications of the purchasable space is screening with the goal of identifying molecules with biological activity without requiring expensive, time-consuming organic syntheses. To study further the drug discovery potential of commercially available molecules, we used the workflow shown in Scheme 1 to selectively remove compounds with unwanted chemical patterns based on advanced filtering rules (Methods and Tables S2−S5). For example, the exclusion of PAINS from chemical libraries prior to testing is of utmost importance in common protein targets, as they represent a major source of false-positive hits in HTS.[45] Remarkably, 13 million molecules did not pass the filters (25.5%), hence posing a threat to unsupervised HTVS. After filtering, this compilation offers a very appealing starting point for the discovery of novel active compounds with desirable medicinal chemistry features.

Partitioning of the Chemical Space into Focused Libraries. The use of focused yet diverse libraries for screening is a common technique in drug discovery.[46] Therefore, we further studied the suitability of the filtered purchasable space in focused library design using semiempirical rules and

DOI: 10.1021/acs.jcim.5b00116
J. Chem. Inf. Model. 2015, 55, 915−924



	MW (Da)	ALOGP	HBD	HBA	RINGS	AROM	ROTB	TPSA (Å^2)	HA
Median purchasable	342.5	2.6	1	4	3	2	6	66.9	24
10th–90th purchasable	*250.3–433.5*	*0.2–4.4*	*0–3*	*2–5*	*1–4*	*1–3*	*3–10*	*37.3–99.3*	*17–30*
Median drug-like	337.4	2.5	1	3	3	2	6	65.4	23
10th–90th drug-like	*249.3–405.1*	*0.2–4.2*	*0–3*	*2–5*	*1–4*	*1–3*	*3–9*	*36.9–95.3*	*17–29*

Figure 2. Physicochemical properties of the purchasable space: absolute frequency distribution histograms (in millions) and statistical analysis of the physicochemical properties of the purchasable and drug-like sets (in sky and dark blue, respectively). Molecular weight (MW); log of the *n*-octanol/water partition coefficient (ALOGP); hydrogen-bond donor (HBD) and acceptor (HBA) groups; number of rings (RINGS), aromatic rings (AROM), and rotatable bonds (ROTB); topological polar surface area (TPSA); and number of heavy atoms (HA) are shown.

predictive methods (Figure 2, dark blue, and Methods). Such partitioning sheds light on the properties, diversity, and coverage of the current purchasable space. 95% of the data set presented drug-like properties (i.e., quantitative estimate of drug-likeness (QED) ≥ 0.35,[27] comprising ~36 million molecules). This amount is noteworthy, as it has been extensively shown that drug-likeness is indicative of the potential success of candidate compounds in the drug discovery pipeline in terms of pharmacokinetics and biological safety.[47] 99.9% of the drug-like compounds additionally pass the classical Lipinki ro5,[6] indicating that the chosen threshold for QED is consistent with the assumptions of the ro5. The collection also comprises over 30 million lead-like chemicals, which are selected based on stricter MW and log $P_{o/w}$ (ALOGP) ranges compared to drug-like compounds, hence possessing bigger optimization potential following the identification of hits.[7]

The physicochemical properties of the drug-like collection are compared to those of the unfiltered set in Figure 2, dark blue. It is worthy to note that the applied chemical filters and drug-likeness selection did not dramatically affect the property distribution parameters. Nonetheless, the removal of large molecules and those containing metals and isotope labeling resulted in a decrease of 5 Da in the median MW, along with a reduction of almost 30 Da on the 90th percentile value to reach 405 Da.

On the basis of commonly used rules, the filtered compounds were additionally classified as fragments and building blocks, NPs, and i-PPIs. The identification of bioactive fragment-like molecules has gained major attention during the last decades and many technologies have been developed aiming at the use of fragments in early stages of drug discovery.[48] For instance, site-directed fragment discovery, or tethering, proved useful in the identification of a potent antagonist of the cytokine IL-2.[49] Filtering the purchasable compounds by means of the "rule of three", as described by Congreve et al., accounted for the identification of over 1.4 million fragment-like commercial chemicals.[28] The adaptation of the "rule of four", described by Morelli et al., after analyzing a manually curated set of i-PPIs,[29] and its application to the commercially available chemical space (Scheme 1) yielded over 3.5 million molecules with i-PPIs-like properties, accounting for around 10% of the entire data set. This is a surprising finding, as this subset contains large, mainly hydrophobic nondrug-like compounds not expected to be extensively represented in commercial catalogs.[50] Finally, we also interrogated the purchasable space for NP-like molecules using the NP-likeness scorer system.[30] By means of this technique, we could identify more than 2.5 million NP-like molecules.

Molecular Diversity Analysis of the Purchasable Space. The chemical diversity and coverage of the generated focused libraries is of crucial importance to determine their suitability and applicability to specific drug discovery campaigns. We initially compared the content of the libraries of i-PPIs-like, drug-like, fragment-like, and NP-like by extracting the overlapping and unique sets of each of them. The resulting Venn diagram, reported in Figure 3, reveals several key aspects regarding the content of the different collections. The drug-like set is richly filled with compounds from the other libraries: It contains practically all fragments and the majority of i-PPIs (68%). It is worthy to note that it comprises a similar amount of molecules from the NP-like and the reference NPs sets (36% and 35%, respectively). Both the i-PPIs- and NP-like collections comprise over one million uniquely represented compounds and, despite the large size and high content of cyclic and polycyclic scaffolds among their molecules, they share less than 1% of compounds. Conversely, most of the fragments and building blocks are part of other data sets, and only 354 compounds could not be assigned to the other groups. Those molecules comprise small, very polar and charged substances with a variety of chemical alarms and reactive groups.

Scaffold Diversity Analysis of the Purchasable Space. Full molecule-based comparisons provide a valid general, yet superficial overview of the data sets: A deeper inspection is required, as exact string matches based on canonical SMILES descriptors of full molecules are not fully informative of the diversity of chemical entities among the data sets. For example, series of analogous compounds, which comprise prevalent structural motifs decorated with a variety of substituents granting diverse physicochemical properties, might not be found during the pairwise comparisons due to the imposed boundaries during filtering.

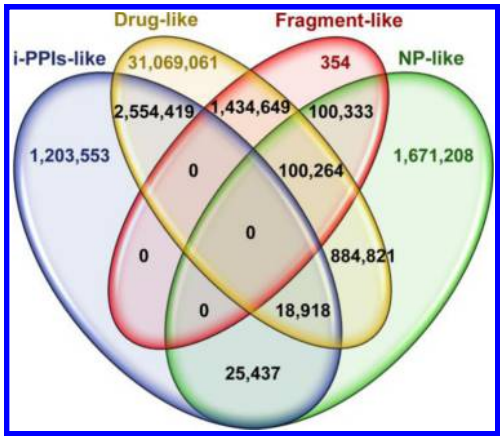


Figure 3. Venn diagram representing the amount of overlapping and unique molecules contained in the focused libraries of i-PPIs-like, drug-like, fragment-like, and NP-like. It is worthy to point out that the overlap between fragments and i-PPIs is forbidden by the mutually exclusive rules applied during library preparation (Scheme 1).

As i-PPIs, fragments, and NPs are of growing interest in HTS, we inspected the scaffold diversity within the related focused libraries following the protocol described in the Methods section. The sets of marketed drugs and reference NPs were included as well for the sake of comparison. Remarkably, we identified thousands of diverse scaffolds in the collection. Some level 0 and level 1 scaffolds appear in a single molecule, i.e. they are singletons, whereas others are substructures of many molecules. For instance, the most common level 0 scaffold within the analyzed collections is benzene. To increase the information content of our study, we mainly considered level 1 scaffolds. Figure 4B represents graphically the overlapping and unique sets of level 1 scaffolds among the focused libraries of i-PPIs-like, fragment-like, and NP-like and the DrugBank and reference NPs sets. Remarkably, each of the focused libraries of the purchasable space contained over 100 000 diverse level 1 scaffolds, 50% of which uniquely represented in each single collection. This finding is particularly noticeable in i-PPIs-like: 162 530 level 1 scaffolds are not represented in any of the other sets (85.3%). An overlap of ~4600 scaffolds between fragments and i-PPIs-like is observed despite the forbidden overlap of their full molecules (Figure 3). Considering that fragment libraries have been successfully applied in the discovery of novel modulators of protein—protein interactions (PPIs),[3,51,52] collected scaffolds within the purchasable space could be analogously used to characterize the binding profile of protein surfaces in therapeutically relevant PPIs.

The analysis in Figure 4B also allows for the comparison of the content of the generated focused libraries with that of the reference NPs and DrugBank sets. From a general perspective, the amount of level 1 scaffolds among the reference sets is substantially smaller than that of the focused libraries (Table 1). The i-PPIs-like contains the largest amount of level 1 scaffolds gathered from the reference NPs (4974), revealing underlying similarities between the molecular cores of those two sets. It also comprises 436 of the level 1 scaffolds in marketed drugs (49.4%). Nonetheless, it should be born in mind that the i-

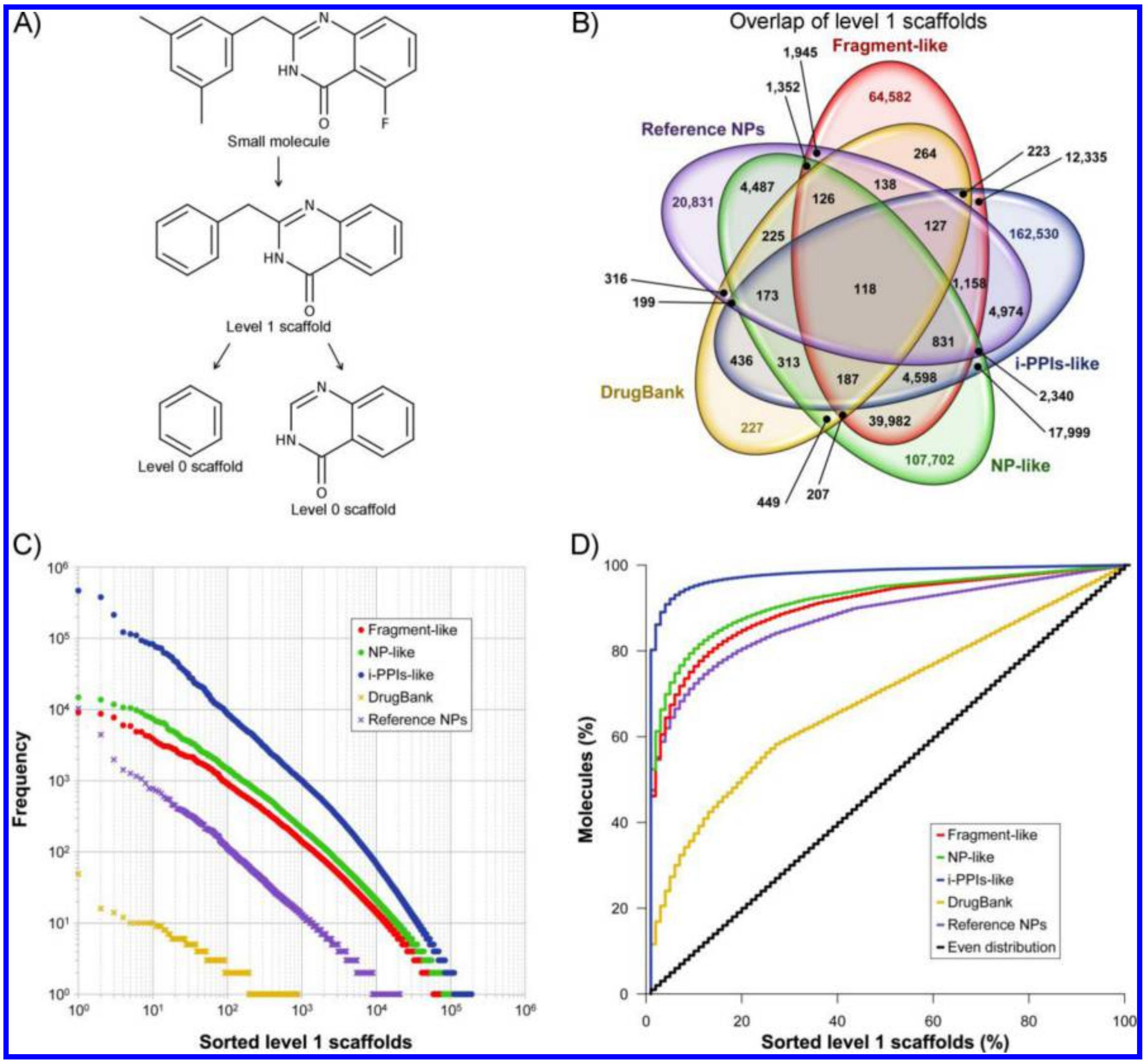


Figure 4. Scaffold diversity analysis of the focused libraries of i-PPIs-like, fragment-like, and NP-like and the reference NPs and DrugBank (see Methods for details). (A) Decomposition of a molecule into level 1 and level 0 scaffold representations. (B) Venn diagram representing the amount of overlapping and unique level 1 scaffolds contained in each collection. (C) Logarithmic scaffold frequency plot. (D) Cumulative scaffold frequency plot.

Table 1. Statistical Results of the Scaffold Diversity Analysis on the Focused Libraries of i-PPIs-like, Fragment-like, and NP-like Sets[a]

data set	M	M_1	N_0	N_1	$N_{1,s}$	M_1/M	N_1/M_1	N_0/N_1	$N_{1,s}/N_1$	P_{50}	P_{90}
i-PPIs-like	3764457	3760841	6118	190643	79379	0.999	0.051	0.032	0.416	0.001	0.035
fragment-like	1435071	1004790	24913	112585	53922	0.700	0.112	0.221	0.479	0.014	0.322
NP-like	2563281	1516720	23876	162783	82763	0.592	0.107	0.147	0.508	0.008	0.262
DrugBank	1493	853	275	882	686	0.571	1.034	0.312	0.778	0.199	0.826
reference NPs	227241	117179	21002	28280	15200	0.516	0.241	0.743	0.537	0.013	0.436

[a] Data for DrugBank and the reference NPs are shown for the sake of comparison. M = number of compounds, M_1 = number of compounds with a level 1 scaffold, N_n = number of level n scaffolds, $N_{1,s}$ = number of singleton level 1 scaffolds, P_n = percentage of level 1 scaffolds represented in n percent of compounds.

PPIs-like set is the largest source of unique level 1 scaffolds among the analyzed collections (190 643, Table 1) despite having the smallest amount of level 0 scaffolds (6118). Scaffolds from the fragment-like collection present the lowest overlap with those of the reference sets: 1945 from the reference NPs and 264 from marketed drugs. The NP-like set, on the other hand, comprises 449 level 1 scaffolds from drugs (50.9%).

Frequency plots of the level 1 scaffolds comprised in the focused libraries of i-PPIs-like, fragment-like, and NP-like are depicted in Figure 4. For comparison, DrugBank compounds and reference NPs are included as well. The plots permit inferring crucial chemical diversity and distribution properties. The linear, scale-free trends observed in the logarithmic scaffold frequency plot (Figure 4C) indicate that the scaffolds are unevenly represented among the collections. For instance, in the i-PPIs-like six scaffolds are represented each in more than 100 000 compounds, whereas thousands of others are singletons. The skewed distribution of scaffolds in molecules has an

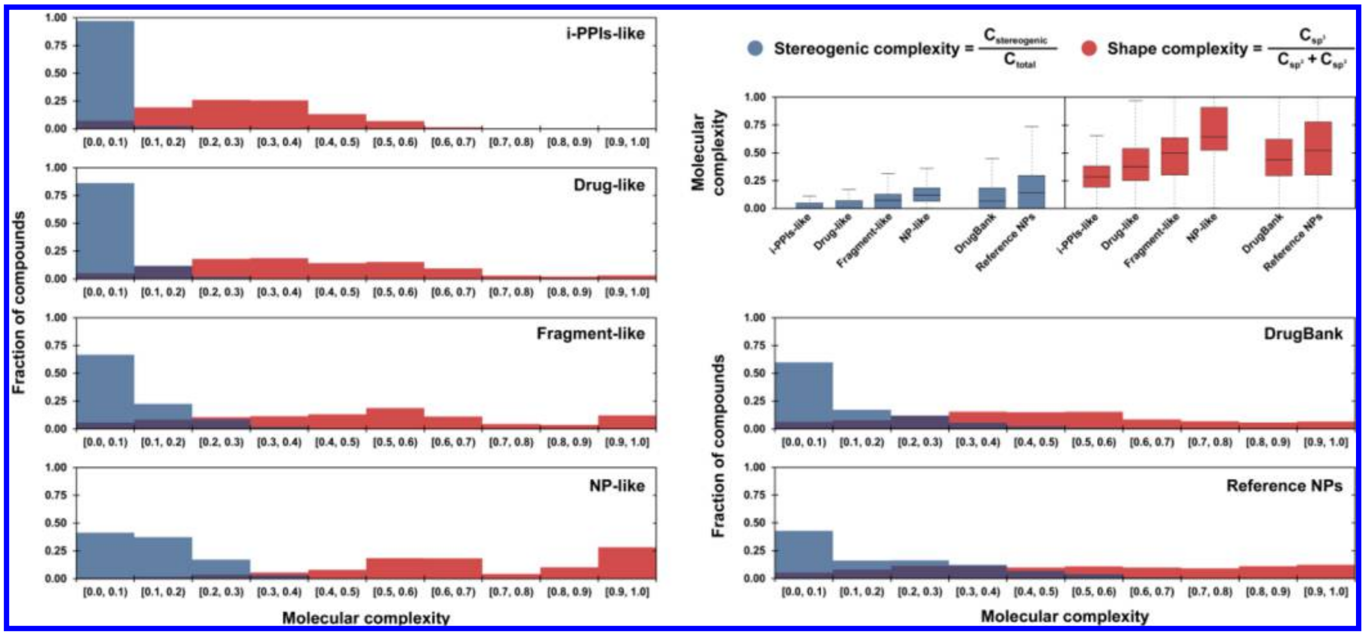


Figure 5. Stereogenic (in blue) and shape (in red) complexity distributions and corresponding boxplot of the generated commercially available focused libraries, DrugBank, and the reference NPs.

impact as well on the cumulative scaffold frequency plot shown in Figure 4D: The rapid saturation of the cumulative curve of level 1 i-PPIs-like scaffolds—one percent of the scaffolds are represented in 80% of the molecules—can be explained by the above-mentioned presence of privileged level 1 scaffolds. The percentage of fragments and NP-like containing one percent of scaffolds is markedly lower −46 and 52%, respectively. The plot reveals an unexpected uneven distribution and certain redundancy of the reference NPs as well −20% of the level 1 scaffolds appear in 80% of the molecules. Remarkably, and despite the substantially larger size of the fragment-like and NP-like data sets compared to the reference NPs, they all present comparable degrees of redundancy.

Table 1 summarizes statistical parameters extracted from the scaffold diversity analysis.[37] For comparison purposes, we considered level 1 (N_1) and level 0 scaffolds (N_0). As commented above, the large collection of level 1 scaffolds extracted from the i-PPIs-like set (190 643) is composed of the smallest compilation of level 0 scaffolds (6118), indicating a low diversity of root chemotypes. Indeed, on average, each level 0 scaffold is represented in 31 level 1 scaffolds (N_1/N_0). This contrasts with the data sets of fragment-like and NP-like, in which each level 0 scaffold is represented in 4.5 and 6.8 level 1 scaffolds, respectively. The ratio is further reduced to 3.2 and 1.3 in marketed drugs and reference NPs, suggesting the existence of a relatively small amount of prevalent successful combinations of level 0 to level 1 scaffolds in those data sets. The diversity within the fragment- and NP-like collections allowed for the identification of more than 23 000 level 0 scaffolds. In terms of level 1 scaffolds (N_1), the data summarized in Table 1 show that the average ratio of scaffolds to molecules (N_1/M_1) in fragment- and NP-like is around 0.1, i.e. each scaffold is represented on average by ∼10 molecules, and the ratio of singletons to scaffolds ($N_{1,s}/N_1$) is around 0.5, i.e. every second scaffold is uniquely represented by a molecule. These values indicate an average higher scaffold diversity for the gathered focused libraries compared to, for instance, those reported by Langdon et al. for a data set of ∼2 million commercially available compounds (N_1 = 81 368, N_1/M_1 = 0.04, $N_{1,s}/N_1$ = 0.77).[37] Consistently with the prevalence of privileged building block combinations in the reference NPs and drugs, the ratio of scaffolds to molecules is ∼4 and ∼1, respectively. This indicates that, on average, each level 1 scaffold from DrugBank constitutes a marketed drug and justifies the growing interest in scaffold diversity analysis and scaffold discovery in drug design. This effect is also reflected in Figure 4D.

We additionally computed the percentage of scaffolds represented in 50% and 90% of compounds (P_{50} and P_{90} respectively, Table 1), which can be extracted from the corresponding cumulative scaffold frequency plot (Figure 4D). These data confirmed the particularly uneven distribution of level 1 scaffolds within the i-PPI-like set: 90% of the molecules (P_{90}) comprise only 3.5% of the scaffolds, whereas in the fragment-like and NP-like sets this number increases to a nonetheless modest 32 and 26%, respectively.

Molecular Complexity Analysis of the Purchasable Space. More than a decade ago it was shown that lead drugs present an intermediate molecular complexity.[25] Since then, structural complexity of small molecules has attracted growing interest in the drug discovery field.[53] For instance, Clemons et al. reported that higher degrees of compound complexity, described either as stereogenic ($C_{stereogenic}/C_{total}$) or shape ($C_{sp3}/[C_{sp2} + C_{sp3}]$) metrics, correlate with higher biological specificity.[23] The content of stereogenic carbon atoms also correlates with other molecular properties, such as the content of HBD and HBA groups, hydrophilicity, and aqueous solubility. By studying the physicochemical and druggability properties of hundreds of protein−drug complexes, we could describe a simple "rule of druggability": druggable, lipophilic binding sites recognize substrates with low stereochemical complexity, whereas drugs with a high content of stereogenic centers bind to low-druggable, polar clefts.[26]

The molecular complexity analysis of the focused libraries of commercially available i-PPIs-like, drug-like, fragment-like, and NP-like, in terms of stereogenicity and hybridization, is reported in Figure 5. In general, we observe that purchasable molecules present an average shape complexity and a low content in stereogenic centers, probably as a result of the majority of companies following more accessible and economic synthetic routes. Moreover, it must be born in mind that most vendors provide their chiral molecules as racemic mixtures. For instance, drug-like molecules present a median stereogenic complexity of 0.00 and a median shape complexity of 0.38. NPs are characterized by a high level of structural complexity: In contrast to the arene-focused synthetic compounds, molecules derived from nature are dominated by chiral, highly hybridized carbons (reference NPs: $\text{median}_{\text{stereogenic complexity}}$ = 0.14, $\text{median}_{\text{shape complexity}}$ = 0.52, Figure 5). Interestingly, we could observe that the collection of commercial NP-like compounds also comprises very complex molecules ($\text{median}_{\text{stereogenic complexity}}$ = 0.12, $\text{median}_{\text{shape complexity}}$ = 0.65). As stereogenic complexity is independent from a molecule's size,[26] it is particularly striking to see that the collection of fragments and building blocks, comprising over 1.4 million molecules, presents a remarkably higher degree of molecular complexity compared to drug-like compounds ($\Delta\text{median}_{\text{stereogenic complexity}}$ = 0.07, $\Delta\text{median}_{\text{shape complexity}}$ = 0.12), indicating that despite the large overlap between those two data sets (Figure 3), fragments are enriched in stereogenic and hybridized carbon atoms. Finally, we also studied the molecular complexity of i-PPIs-like molecules (Figure 5). PPIs present physicochemical properties significantly distinct from those observed in common target classes, thus their modulation by small molecules remains challenging.[50] They are dominated by large, flat, mainly hydrophobic surfaces with scarce presence of interaction features. Accordingly, commercially available i-PPIs-like are represented by the simplest molecules studied herein ($\text{median}_{\text{stereogenic complexity}}$ = 0.00, $\text{median}_{\text{shape complexity}}$ = 0.29).

Taken together, these findings indicate that both NPs and fragments might prove particularly useful in the identification of privileged scaffolds binding to low-druggable proteins, which recognize complex compounds, whereas i-PPIs-like can be used to characterize lipophilic recognition pockets or protein surfaces.[26] Bioactive fragments identified during PPI characterization screening could be afterward enlarged taking into consideration an overall decrease of complexity to cope with a target's requirements. The boxplot in Figure 5 also shows that not only compounds comprised in those focused libraries cover a wide range of stereogenic and shape complexities, but they also follow the tendency: i-PPIs-like < drug-like < fragment-like < NP-like. The comparison of the purchasable sets with the reference NPs and DrugBank shows that, all in all, they present similar molecular complexities.

Molecular Complexity Plays a (Hidden) Major Role in Library Design. It has been reported that compounds contained in commercial catalogs are structurally simple.[54] Likewise, diversity and complexity of fragment libraries are currently under discussion.[55,56] Nonetheless, the purchasable space collected herein comprises highly diverse scaffolds represented by molecules ranging from very simple, large, planar i-PPIs-like to richly stereogenic and hybridized NP-like (Figures 4 and 5). Our findings specifically indicate that the current commercial fragment-like space comprises compounds that, despite their size, encompass a degree of molecular complexity higher than that of drug-like compounds and comparable to that of marketed drugs. Noteworthy, and due to their small size (median heavy atoms $(\text{HAs})_{\text{fragment-like}}$ = 16, median $\text{HAs}_{\text{drug-like}}$ = 23), the chance of identifying hit compounds using fragments increases compared to full-sized ligands.[53,57] Furthermore, the herein discussed structural simplicity of i-PPIs-like has implications in the therapeutic modulation of protein interfaces: on the one hand, it suggests that planar, synthetically accessible commercial compounds can be an approach to accelerated probing of PPIs. On the other hand, it poses structural complexity as yet another molecular descriptor that differentiates i-PPIs-like from other compounds, hence it could be used to prefilter large data sets.

Free Availability of the Collected Commercially Available Compounds. The presented analyses demonstrate that the collected data set of commercially available compounds can be of huge benefit for drug discovery, as it surely contains bioactive agents. In an attempt to provide the community with the compounds, we requested free accessibility to the catalogs of the collaborating companies. In total, we received the agreement from 49 companies (Table S1), amounting over 37 million unique compounds. This unfiltered compilation, termed PurchasableBoX, is freely available as a simple SMILES-formatted flat file under http://pbox.pharmaceutical-bioinformatics.org. The content of the PurchasableBoX and its partitioned focused libraries was compared to that of the complete set studied herein and the ZINC database (Figure 6).

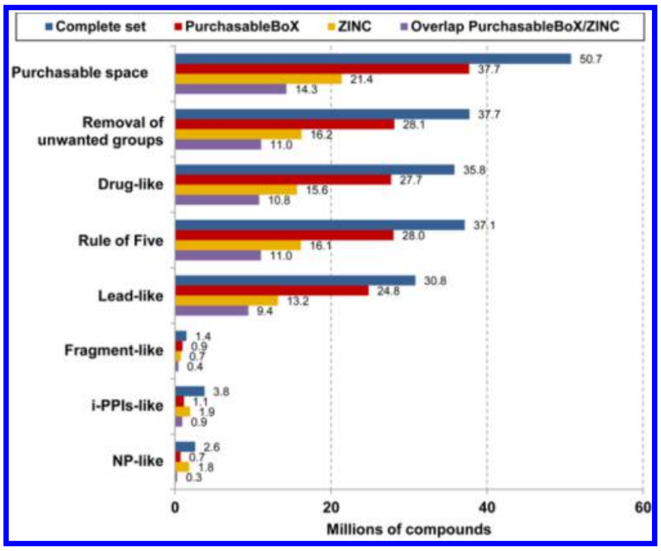


Figure 6. Content of the unfiltered, freely available data set PurchasableBoX and its partitioned focused libraries compared to that of the current study and the ZINC database.[16] PurchasableBoX is freely available under http://pbox.pharmaceutical-bioinformatics.org.

The collection comprises 74% of the in-house purchasable space and is complementary to ZINC: PurchasableBoX contains 23.4 million molecules not found in ZINC, and vice versa, ZINC contains 7.1 million molecules not found in PurchasableBoX. Importantly, it is enriched in drug-like (over 27 million), fragment-like, i-PPIs-like, and NP-like compounds.

■ CONCLUSIONS

In the present manuscript we have collected ~125 million commercially available compounds, comprising over 68 million unique chemicals, and analyzed their physicochemical properties and scaffold diversity. This collection contained over 90% of marketed drugs and over 80 000 referenced NPs (36%). The application of advanced medicinal chemistry filters, including PAINS-based selectors (complete lists included as Supporting

Information), revealed over 13 million molecules that did not fulfill substructure patterns; hence the unsupervised use of commercial catalogs for screening is discouraged.

The partition of the purchasable space into drug discovery relevant focused libraries, including drug-like, i-PPIs-like, fragment-like, and NP-like, is a common approach to bias screening sets. It enabled the identification of overlapping spaces among the libraries: for example, most fragments presented additionally drug-like properties, yet only a residual overlap was detected between i-PPIs-like and NP-like. A detailed scaffold diversity analysis revealed thousands of scaffolds in commercially available compounds unique for each data set. Their scaffold distributions were compared to those of a reference NPs set and the marketed drugs from DrugBank, showing characteristic behaviors: In DrugBank, scaffolds are distributed evenly among the molecules, whereas in the i-PPIs-like library one percent of the scaffolds is present in 80% of the molecules. By comparing the ratio of level 0 to level 1 scaffolds, we could also observe distinct patterns among the libraries: i-PPIs-like presents a large combination of building blocks, whereas NPs exploits only a limited number of potential chemotypes.

An analysis of the stereogenic and shape complexity distributions of compounds comprised in those data sets additionally revealed that they cover a wide range of molecular complexity, from very simple i-PPIs-like compounds to complex NP-like, following the tendency: i-PPIs-like < drug-like < fragment-like < NP-like.

As the collection studied herein can be of huge interest for drug discovery, we offer over 37 million purchasable, unfiltered molecules for which we have received the agreement from the vendors for publishing under: http://pbox.pharmaceutical-bioinformatics.org. This compilation can be complementarily used with ZINC to build large, comprehensive sets for screening or the preparation of focused libraries. It also comprises thousands of NPs ready for delivery.

ASSOCIATED CONTENT

Supporting Information

Table S1 contains a complete listing of vendors. Applied classical and advanced chemical filters are detailed in Tables S2, S3, S4, and S5. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00116.

AUTHOR INFORMATION

Corresponding Author

*E-mail: stefan.guenther@pharmazie.uni-freiburg.de.

Author Contributions

X.L. conceived the study, carried out the data collection and the analyses, and wrote the manuscript. B.A.G. participated in the design of the study and, together with S.B., provided computational assistance and implemented some of the applied tools. S.G. participated in the design of the study, coordinated the project, and helped to draft the manuscript. All authors have read and given approval to the final version of the manuscript.

Funding

This work has received funding from the Excellence Initiative of the German Federal and State Governments through the Junior Research Group Program (ZUK 43). The work from X.L. has been funded by the Federal State of Baden-Württemberg (Ministry of Arts and Science) through the "Zukunftsoffensive IV (ZO IV) Juniorprofessoren-Programm" and by the DFG (1225/3-1).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Kiran K. Telukunta for technical support. We kindly thank the supplier companies for providing their catalogs of small molecules free of charge, especially those supporting open access to their catalog.

ABBREVIATIONS

ALOGP, predicted additive octanol/water partition coefficient; AROM, number of aromatic rings; CAS, Chemical Abstracts Service; HA, number of heavy atoms; HBA, number of hydrogen-bond acceptor groups; HBD, number of hydrogen-bond donor groups; HTS, high-throughput screening; i-PPI, inhibitor of protein−protein interaction; MW, molecular weight; NP, natural product; PAINS, pan assay interference compounds; PPI, protein−protein interaction; TPSA, topological polar surface area; QED, quantitative estimate of drug-likeness; RINGS, number of rings; ROTB, number of rotatable bonds; ro5, rule of five

REFERENCES

(1) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(2) Sukumar, N.; Das, S. Current Trends in Virtual High Throughput Screening Using Ligand-Based and Structure-Based Methods. *Comb. Chem. High Throughput Screen.* **2011**, *14*, 872−888.

(3) Lucas, X.; Wohlwend, D.; Hügle, M.; Schmidtkunz, K.; Gerhardt, S.; Schüle, R.; Jung, M.; Einsle, O.; Günther, S. 4-Acyl Pyrroles: Mimicking Acetylated Lysines in Histone Code Reading. *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 14055−14059.

(4) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188−195.

(5) Harvey, A. L. Natural Products in Drug Discovery. *Drug Discovery Today* **2008**, *13*, 894−901.

(6) Lipinski, C. A. Drug-Like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(7) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−263.

(8) Zhou, J. Z. Chemoinformatics and Library Design. *Methods Mol. Biol.* **2011**, *685*, 27−52.

(9) Dolle, R. E. Historical Overview of Chemical Library Design. *Methods Mol. Biol.* **2011**, *685*, 3−25.

(10) Cragg, G. M.; Newman, D. J. Natural Products: A Continuing Source of Novel Drug Leads. *Biochim. Biophys. Acta* **2013**, *1830*, 3670−3695.

(11) Patel, R. N. Chemo-Enzymatic Synthesis of Pharmaceutical Intermediates. *Expert Opin. Drug Discovery* **2008**, 3, 187−245.

(12) Hou, J.; Liu, X.; Shen, J.; Zhao, G.; Wang, P. G. The Impact of Click Chemistry in Medicinal Chemistry. *Expert Opin. Drug Discovery* **2012**, *7*, 489−501.

(13) Gantt, R. W.; Peltier-Pain, P.; Thorson, J. S. Enzymatic Methods for Glyco(Diversification/Randomization) of Drugs and Small Molecules. *Nat. Prod. Rep.* **2011**, *28*, 1811−1853.

(14) Kodadek, T. The Rise, Fall and Reinvention of Combinatorial Chemistry. *Chem. Commun. (Camb.)* **2011**, *47*, 9757−9763.

(15) Pollier, J.; Moses, T.; Goossens, A. Combinatorial Biosynthesis in Plants: A (P)Review on Its Potential and Future Exploitation. *Nat. Prod. Rep.* **2011**, *28*, 1897−1916.

(16) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. Zinc: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(17) Daga, P. R.; Polgar, W. E.; Zaveri, N. T. Structure-Based Virtual Screening of the Nociceptin Receptor: Hybrid Docking and Shape-Based Approaches for Improved Hit Identification. *J. Chem. Inf. Model.* **2014**, *54*, 2732−2743.

(18) Lucas, X.; Simon, S.; Schubert, R.; Günther, S. Discovery of the Inhibitory Effect of a Phosphatidylinositol Derivative on P-Glycoprotein by Virtual Screening Followed by in Vitro Cellular Studies. *PLoS One* **2013**, *8*, e60679.

(19) Parmenopoulou, V.; Kantsadi, A. L.; Tsirkone, V. G.; Chatzileontiadou, D. S.; Manta, S.; Zographos, S. E.; Molfeta, C.; Archontis, G.; Agius, L.; Hayes, J. M.; Leonidas, D. D.; Komiotis, D. Structure Based Inhibitor Design Targeting Glycogen Phosphorylase B. Virtual Screening, Synthesis, Biochemical and Biological Assessment of Novel N-Acyl-Beta-D-Glucopyranosylamines. *Bioorg. Med. Chem.* **2014**, *22*, 4810−4825.

(20) Chen, C.; Wang, T.; Wu, F.; Huang, W.; He, G.; Ouyang, L.; Xiang, M.; Peng, C.; Jiang, Q. Combining Structure-Based Pharmacophore Modeling, Virtual Screening, and in Silico Admet Analysis to Discover Novel Tetrahydro-Quinoline Based Pyruvate Kinase Isozyme M2 Activators with Antitumor Activity. *Drug Des. Devel. Ther.* **2014**, *8*, 1195−1210.

(21) ChemicalToolBoX. https://github.com/bgruening/galaxytools/tree/master/chemicaltoolbox (accessed February 10, 2015).

(22) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(23) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U S A* **2010**, *107*, 18787−18792.

(24) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52*, 6752−6756.

(25) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.

(26) Lucas, X.; Günther, S. Using Chiral Molecules as an Approach to Address Low-Druggability Recognition Sites. *J. Comput. Chem.* **2014**, *35*, 2114−2121.

(27) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90−98.

(28) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

(29) Morelli, X.; Bourgeas, R.; Roche, P. Chemical and Structural Lessons from Recent Successes in Protein-Protein Interaction Inhibition (2P2I). *Curr. Opin. Chem. Biol.* **2011**, *15*, 475−481.

(30) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68−74.

(31) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33.

(32) Chen, C. Y. Tcm Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in Silico. *PLoS One* **2011**, *6*, e15939.

(33) Chapman and Hall Dictionary of Natural Products. http://dnp.chemnetbase.com (accessed September 15, 2008).

(34) Lucas, X.; Senger, C.; Erxleben, A.; Grüning, B. A.; Döring, K.; Mosch, J.; Flemming, S.; Gunther, S. StreptomeDB: A Resource for Natural Compounds Isolated from *Streptomyces* Species. *Nucleic Acids Res.* **2013**, *41*, D1130−1136.

(35) Saubern, S.; Guha, R.; Baell, J. B. Knime Workflow to Assess Pains Filters in Smarts Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol. Inf.* **2011**, *30*, 847−850.

(36) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A Comprehensive Resource for 'Omics' Research on Drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−1041.

(37) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51*, 2174−2185.

(38) Chemical Abstracts Service (CAS). http://www.cas.org (accessed December 10, 2014).

(39) iResearch Library. www.chemnavigator.com/cnc/products/iRL.asp (accessed December 10, 2014).

(40) Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged Scaffolds for Library Design and Drug Discovery. *Curr. Opin. Chem. Biol.* **2010**, *14*, 347−361.

(41) Boguski, M. S.; Mandl, K. D.; Sukhatme, V. P. Drug Discovery. Repurposing with a Difference. *Science* **2009**, *324*, 1394−1395.

(42) Gupta, S. C.; Sung, B.; Prasad, S.; Webb, L. J.; Aggarwal, B. B. Cancer Drug Discovery by Repurposing: Teaching New Tricks to Old Dogs. *Trends Pharmacol. Sci.* **2013**, *34*, 508−517.

(43) Proudfoot, J. R. The Evolution of Synthetic Oral Drug Properties. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1087−1090.

(44) Zuegg, J.; Cooper, M. A. Drug-Likeness and Increased Hydrophobicity of Commercially Available Compound Libraries for Drug Screening. *Curr. Top. Med. Chem.* **2012**, *12*, 1500−1513.

(45) Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481−483.

(46) Gregori-Puigjane, E.; Mestres, J. Coverage and Bias in Chemical Library Design. *Curr. Opin. Chem. Biol.* **2008**, *12*, 359−365.

(47) Yusof, I.; Segall, M. D. Considering the Impact Drug-Like Properties Have on the Chance of Success. *Drug Discovery Today* **2013**, *18*, 659−666.

(48) Scott, D. E.; Coyne, A. G.; Hudson, S. A.; Abell, C. Fragment-Based Approaches in Drug Discovery and Chemical Biology. *Biochemistry* **2012**, *51*, 4990−5003.

(49) Braisted, A. C.; Oslob, J. D.; Delano, W. L.; Hyde, J.; McDowell, R. S.; Waal, N.; Yu, C.; Arkin, M. R.; Raimundo, B. C. Discovery of a Potent Small Molecule IL-2 Inhibitor through Fragment Assembly. *J. Am. Chem. Soc.* **2003**, *125*, 3714−3715.

(50) Bourgeas, R.; Basse, M. J.; Morelli, X.; Roche, P. Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. *PLoS One* **2010**, *5*, e9598.

(51) Hajduk, P. J.; Greer, J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211−219.

(52) Bower, J. F.; Pannifer, A. Using Fragment-Based Technologies to Target Protein-Protein Interactions. *Curr. Pharm. Des.* **2012**, *18*, 4685−4696.

(53) Leach, A. R.; Hann, M. M. Molecular Complexity and Fragment-Based Drug Discovery: Ten Years On. *Curr. Opin. Chem. Biol.* **2011**, *15*, 489−496.

(54) Dandapani, S.; Marcaurelle, L. A. Grand Challenge Commentary: Accessing New Chemical Space for 'Undruggable' Targets. *Nat. Chem. Biol.* **2010**, *6*, 861−863.

(55) Koster, H.; Craan, T.; Brass, S.; Herhaus, C.; Zentgraf, M.; Neumann, L.; Heine, A.; Klebe, G. A Small Nonrule of 3 Compatible Fragment Library Provides High Hit Rate of Endothiapepsin Crystal Structures with Various Fragment Chemotypes. *J. Med. Chem.* **2011**, *54*, 7784−7796.

(56) Wilde, F.; Link, A. Advances in the Design of a Multipurpose Fragment Screening Library. *Expert Opin. Drug Discovery* **2013**, *8*, 597−606.

(57) Bembenek, S. D.; Tounge, B. A.; Reynolds, C. H. Ligand Efficiency and Fragment-Based Drug Discovery. *Drug Discovery Today* **2009**, *14*, 278−283.