JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

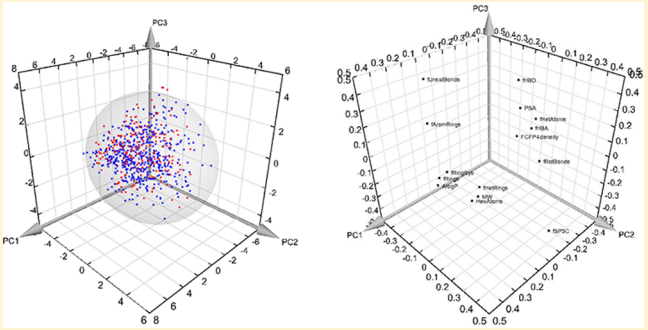
Locating Sweet Spots for Screening Hits and Evaluating Pan-Assay Interference Filters from the Performance Analysis of Two Lead-like Libraries

N. Yi Mok,[†,‡] Sara Maxe,[†] and Ruth Brenk*[,§]

Drug Discovery Unit, Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, U.K.

Supporting Information

ABSTRACT: The efficiency of automated compound screening is heavily influenced by the design and the quality of the screening libraries used. We recently reported on the assembly of one diverse and one target-focused lead-like screening library. Using data from 15 enzyme-based screenings conducted using these libraries, their performance was investigated. Both libraries delivered screening hits across a range of targets, with the hits distributed across the entire chemical space represented by both libraries. On closer inspection, however, hit distribution was uneven across the chemical space, with enrichments observed in octants characterized by compounds at the higher end of the molecular weight and lipophilicity spectrum for lead-like compounds, while polar and sp^3-carbon atom rich compounds were underrepresented among the screening hits. Based on these observations, we propose that screening libraries should not be evenly distributed in lead-like chemical space but be enriched in polar, aliphatic compounds. In conjunction with variable concentration screening, this could lead to more balanced hit rates across the chemical space and screening hits of higher ligand efficiency will be captured. Apart from chemical diversity, both screening libraries were shown to be clean from any pan-assay interference (PAINS) behavior. Even though some compounds were flagged to contain PAINS structural motifs, some of these motifs were demonstrated to be less problematic than previously suggested. To maximize the diversity of the chemical space sampled in a screening campaign, we therefore consider it justifiable to retain compounds containing PAINS structural motifs that were apparently clean in this analysis when assembling screening libraries.



INTRODUCTION

The advent of high-throughput screening (HTS) for drug discovery in the 1980s enabled the rapid screening of diverse chemical compounds during hit identification. In the early days of HTS, compound collections were mainly assembled from internal resources and often also contained compounds from previous company activities such as dyes and fine chemicals.[1] Through the introduction of combinatorial chemistry, the size of screening libraries expanded. However, their quality could remain poor owing to limited compound diversity and undesirable compound properties.[1] Since the establishment of drug-like and lead-like concepts into drug discovery,[2−4] physicochemical properties represented key parameters for compound selection that led to higher quality hits in HTS campaigns.[1] Nowadays, the focus of enhancing screening libraries lies on increasing scaffold diversity for general screening purposes and the assembly of target- or gene-family tailored libraries, often under consideration of physicochemical property constraints to maintain the lead-like character of the selected compounds.[1,5,6]

Lead-like compounds are considered to be smaller and structurally less complex than drug-like molecules.[3,4] This allows expansion of molecules in lead optimization, and at the same time enables more efficient sampling of chemical space since the latter is estimated to expand exponentially for every extra heavy atom in a molecule.[7] Moreover, compound affinities of lead-like molecules can be detected at typical screening concentrations in the low micromolar range using conventional HTS laboratory setup without the need for sensitive detection methodologies such as biophysical techniques commonly required for fragment-based hit discovery.[8]

A common problem associated with compound libraries is the presence of compounds displaying promiscuous behavior. These false positives, sometimes also termed "frequent hitters" or more recently, pan-assay interference compounds (PAINS),[9] display nonspecific enzyme inhibition through mechanisms including, but not limited to, compound aggregation and covalent protein binding through the presence of reactive functional groups (see the work of Thorne et al. for a comprehensive review).[10] Compounds displaying promiscuous behavior reduce the efficiency of hit identification in library

Received: August 15, 2012
Published: March 4, 2013

screening by wasting valuable resources in attempted, but unsuccessful, compound optimization efforts. Several recent publications have highlighted structural characteristics and motifs responsible for such behavior that the authors suggested as useful filters in screening libraries to enhance the efficiency of hit discovery.[9,11,12]

Table 1. Lead-like Selection Criteria Used for Compounds in DSL and FKL[13]

selection criteria	definition
size and physicochemical properties	10−27 heavy atoms
	<4 hydrogen-bond donors
	<7 hydrogen-bond acceptors
	0 < (hydrogen-bond donors + hydrogen-bond acceptors) < 10
	$0 \leq$ ClogP/ClogD ≤ 4
	at least one nonring atom if compound contains only one ring system
limited complexity	<8 rotatable bonds
	<5 ring systems
	no ring systems with more than two fused rings
absence of unwanted functionalities	exclusion of compounds containing potentially reactive, metabolically labile or toxic groups (as defined in the work of Brenk et al.[13])

At the University of Dundee, we have reported the assembly of several screening libraries, including a diverse screening library (DSL) and target-focused libraries against kinases (FKL) and ion channels, all compiled using physicochemical properties compliant to lead-like criteria (Table 1).[13,14] To date, a number of enzyme- and cell-based screens have been carried out using these libraries, with a wide target spectrum across multiple species of organisms using various assay readout technologies. These screening results provide a valuable opportunity to assess the performance of lead-like screening libraries. In the current work, we report the analysis of results collected from 15 enzyme-based screenings conducted using DSL and FKL. We evaluated the utilization of chemical space represented by each library and the distribution of screening hits within this chemical space. We then assessed whether any library compounds should be classified as pan-assay interference (PAINS) according to the definitions of Baell and Holloway. Finally, we investigated if compounds containing previously identified structural motifs of PAINS were indeed promiscuous inhibitors in our screens.[9] On the basis of these analyses, we give recommendations on the composition of lead-like libraries and associated screening practice to obtain an even distribution of hit compounds in the chemical space represented, and the application of PAINS filters to remove compounds when assembling screening libraries.

RESULTS

Data Collection. The data from 15 enzyme-based screening campaigns were selected for analysis (Tables 2 and 3).[15−29] The applied assays were end point assays using a variety of readout technologies, with typical compound concentration at 30 μM. All campaigns discussed in this analysis had Z' values >0.5, indicating excellent assay performance.[30] To allow consistent comparison across multiple assays, only compounds that have been screened against all targets were included in the analysis. This led to a collection of 59 443 compounds from DSL against seven targets and 3287 compounds from FKL against 10 targets, which together represented data from five different assay readout technologies (Tables 2 and 3). Primary hits were defined as compounds above a certain threshold percentage inhibition value that was derived from the mean percentage inhibition value and its standard deviation for each individual assay. Compounds interfering with the particular assay readout technology, for example, colored compounds in colorimetric assays, were excluded. Followed-up hits were defined as primary hits that subsequently had identity and purity confirmed using LC-MS, IC_{50} values determined (a minimum of two independent measurements), and a Hill slope of the log concentration−response curve within the range 0.7−1.5. The latter criterion was applied to only include inhibitors that were potentially competitive with respect to the substrate and to exclude promiscuous inhibitors due to aggregate formation that often result in high Hill slopes.[31] It is noteworthy that the number of primary hits selected for subsequent IC_{50} determination was dependent on the capacity of the individual biological assay and the presence of structure−activity relationships within the primary screening data. Therefore, not every compound was followed up in certain assays, particularly those which resulted in a large number of primary hits. Hence, one should not draw conclusions about false positives based on the difference in the number of compounds between the two stages. In total, DSL delivered 1720 primary hits and 302 followed-up hits, whereas FKL delivered 747 primary hits and 255 followed-up hits (Tables 2 and 3).

Hit Compound Distribution in Chemical Space. The chemical space represented by each screening library was defined using 15 descriptors characterizing the physicochemical properties and molecular complexity of the screening compounds (Table 4). These descriptors are mainly common

Table 2. Number of Compounds Reported As Primary Hits and Followed-up Hits, Together with the Hit Rates and Readout Technology Used, for Each Biological Target Screened Using DSL

target	target class	no. of primary hits (hit rate (%))	no. of followed-up hits	readout technology
*Hs*OGA[28]	glycosidase	38 (0.06)	6	fluorescence
picornaviral 3C cysteine protease[20]	cysteine protease	3 (0.005)	0	fluorescence
*Tb*NMT[15,18]	acyltransferase	275 (0.46)	111	scintillation proximity
*Hs*OGT[27]	glycosyltransferase	132 (0.22)	10	scintillation proximity
*Tb*TryS[29]	ligase	611 (1.03)	127	colorimetric
*Tb*TryR[17]	oxidoreductase	722 (1.21)	51	colorimetric
*Tb*UAP[24]	nucleotidyltransferase	7 (0.01)	3	colorimetric
total		**1720**[a] **(2.9)**[b]	**302**[a]	

[a]After removing duplicate compounds. [b]Percentage of compounds that were active in at least one screen.

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

Table 3. Number of Compounds Reported As Primary Hits and Followed-up Hits, Together with the Hit Rates and Readout Technology Used, for Each Biological Target Screened Using FKL

target	target class	no. of primary hits (hit rate (%))	no. of followed-up hits	readout technology
*Hs*OGT[27]	glycosyltransferase	5 (0.15)	1	scintillation proximity
*Tb*TryS[29]	ligase	25 (0.76)	19	colorimetric
*Bp*HSP90[21]	ATP-dependent chaperone	14 (0.43)	1	fluorescence polarization
*Lm*CRK3[16]	Ser/Thr kinase	72 (2.19)	45	fluorescence polarization
*Pf*CDPK5[25]	Ser/Thr kinase	43 (1.31)	20	fluorescence
*Tb*PLK[22]	Ser/Thr kinase	62 (1.89)	6	luminescence
*Tb*GSK3[23]	Ser/Thr kinase	406 (12.4)	55	luminescence
*Tb*PK53[26]	Ser/Thr kinase	199 (6.05)	62	luminescence
*Tb*PK50[26]	Ser/Thr kinase	425 (12.9)	82	luminescence
*Ec*IspE[19]	GHMP kinase	1 (0.03)	1	luminescence
total		**747**[a] **(22.7)**[b]	**255**[a]	

[a]After removing duplicate compounds. [b]Percentage of compounds that were active in at least one screen.

Table 4. Descriptors Used for Describing the Chemical Space Represented by Each Screening Library

descriptor	abbreviation
molecular weight	MW
number of heavy atoms	HevAtoms
logarithmic octanol/water partition coefficient	ALogP
polar surface area	PSA
fraction of [a]	
hydrogen-bond acceptors	fHBA
hydrogen-bond donors	fHBD
heteroatoms	fHetAtoms
rotatable bonds	fRotBonds
unsaturated bonds	fUnsatBonds
rings	fRings
heterocycles	fHetRings
aromatic rings	fAromRings
ring systems	fRingSys
sp^3-hybridized carbon atoms [b]	fSP3C
normalized functional class extended connectivity fingerprints[32] [a]	FCFP4density

[a]Normalized relative to the number of heavy atoms unless stated otherwise. [b]Normalized relative to the number of carbon atoms.[33]

parameters used for describing molecular features and binding capabilities of small molecules.[32,33] All categorical descriptors with discrete unit values were normalized relative to the number of heavy atoms or the number of carbon atoms to reflect the intrinsic trends of each descriptor independent of the size of a molecule.

Principal component analysis (PCA) was performed on the descriptor matrix to visualize the chemical space represented by each screening library (Figure 1). For DSL, the first three principal components accounted for 22%, 20%, and 16% of the X-variance, respectively, with a cumulative R^2 of 0.58 (Figure 1a and b). The mapping of hit compounds in the projected chemical space suggested that all the primary hits and followed-up hits were distributed across the entire chemical space, with no particular regions observed where no screening hits were reported. Similarly, the mapping of hit compounds in the 3D PCA projection for FKL (cumulative R^2 = 0.62, Figure 1c and d) displayed a scattered distribution of all the primary hits and followed-up hits across the entire chemical space. Again, there were no particular regions of the chemical space where no screening hits were reported.

In an attempt to quantify the distribution of primary hits and followed-up hits in the screening libraries, the volume of chemical space represented in the 3D PCA plots was divided into eight regions (octants) around the center of origin (Figure 2a). The percentage of each category of compounds in all eight octants of the PCA plots was then assessed (Figure 2b and c).

The compounds in DSL were evenly distributed across all eight octants, with 10−15% of compounds in each octant (Figure 2b). Each octant also contained primary hits and followed-up hits from a broad range of targets (Figures S1 and S2, Supporting Information). However, the hit rates per octant varied. Of notable differences were octants 1 and 2, where approximately a 1.5-fold enrichment of primary hits and follow-up hits relative to the percentage of all screening compounds in the particular octant was observed. Mapping of descriptors in these octants on the loading plot (Table 5) revealed that these regions of chemical space were characterized by aromatics (octant 1) and heavy, lipophilic compounds (octant 2). The average molecular weight of compounds within these octants was, respectively, 21 and 45 Da higher than the average of the full library (318 Da), whereas the average ALogP was increased by 0.6−0.8 units compared to the DSL average (2.6) (Figure 3a). On the contrary, octants 4 and 8 displayed a 2-fold decrease in the percentage of primary hits and followed-up hits as compared to the percentage of all screening compounds (Figure 2b). These regions of chemical space featured more polar and heteroatom rich compounds (PSA = 113 (octant 4) vs 77 Å^2 for the DSL average; fHetAtoms 20% and 32% above the DSL average, respectively), compounds with higher fraction of heterocycles (fHetRings 23% and 35% above the DSL average, respectively), and compounds with higher FCFP4density (FCFP4density 7% and 13% above the DSL average, respectively) (Figure 3a). A decrease in the percentage of primary hits and followed-up hits was also observed in octant 7 (Figure 2b), where compounds were characterized by a high fraction of sp^3-carbon atoms (fSP3C 0.49, 88% above the DSL average, Figure 3a).

For FKL, the entire library was again evenly distributed across all eight octants, with each comprising of 10−15% of screening compounds (Figure 2c). Again, all octants contained primary hits and followed-up hits from a range of targets (Figures S1 and S2, Supporting Information). A similar trend as in DSL was observed with an enrichment of primary hits and followed-up hits in octants 1 (1.4-fold increase) and 2 (2-fold increase) where the chemical space was characterized by heavy, lipophilic compounds (octant 1) and aromatics (octant 2) (Table 5). For instance, the average molecular weight of compounds in octant 1 was 70 Da higher than the average of

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

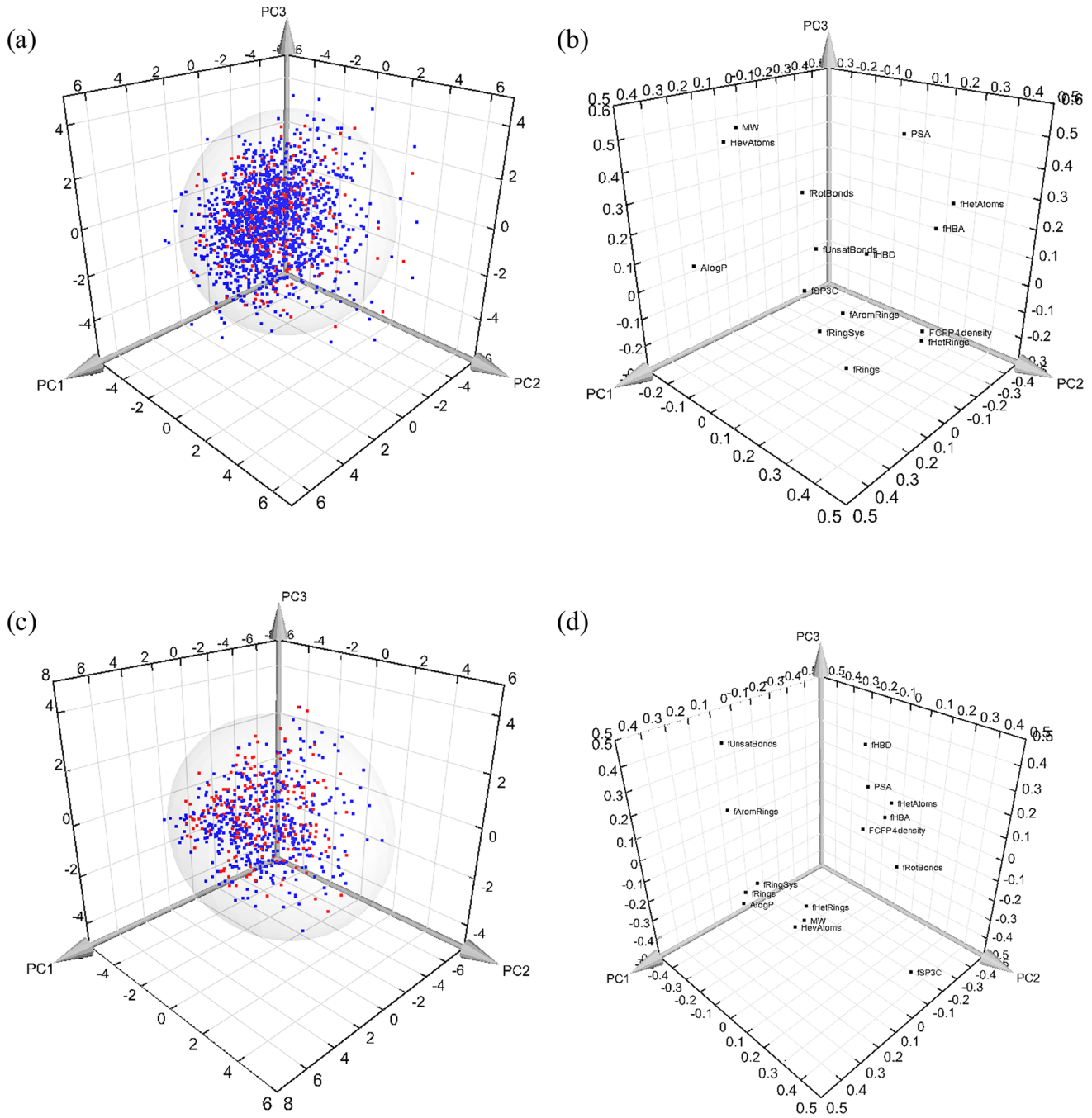


Figure 1. Scoring plots (left) and corresponding loading plots (right) of the PCA of the chemical space represented by DSL (a and b) and FKL (c and d). The gray ellipsoid corresponds to a confidence level of 95% of Hotelling's T^2 distribution. Primary hits are colored in blue, and followed-up hits are colored in red.

the full library (318 Da), and the average ALogP was 1.1 units higher than the FKL average (2.7) (Figure 3b). The regions of chemical space which had a decrease in percentage of primary hits and followed-up hits were octants 4 (3-fold decrease) and 8 (3.2-fold decrease) (Figure 2c), where the chemical space was characterized by polar compounds (octant 4; PSA = 93 vs 72 Å^2 for the FKL average) and aliphatic compounds (octant 8; fSP3C 0.42, 99% above the FKL average, Figure 3b).

We then proceeded to further evaluate the average ligand efficiency of followed-up hits within each octant (Figure 4).[34] As expected, hits with the highest average ligand efficiency were located in octants characterized by compounds with the lowest average molecular weight (octant 8 for DSL and octants 3 and 7 for FKL). However, we also observed differences in the average ligand efficiency in octants where the average molecular weight of followed-up hits was comparable. For instance, octants 1−4 of DSL contained followed-up hits of similar size (MW = 360−372 Da, Figure 4a). Out of those hits, the polar and heteroatom rich compounds in octant 4 (Figure 3a) displayed the highest average ligand efficiency (0.30 kcal mol^{-1} per heavy atom). This trend was also present in FKL. Compounds in octant 4 (polar compounds, Figure 3b) achieved the highest average ligand efficiency (0.36 kcal mol^{-1} per heavy atom) among the octants containing

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

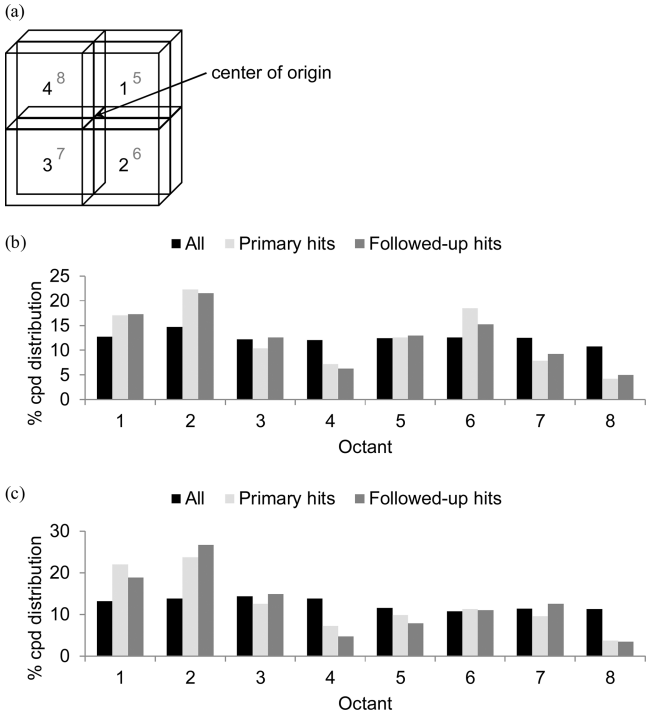


Figure 2. (a) Illustrative diagram of octant assignments of the PCA diagrams in Figure 1. The center of origin (0, 0, 0) is at the intersection of all eight octants in the middle. (b and c) Percentage distribution of the full library, primary hits, and followed-up hits in each of the eight octants in DSL (b) and FKL (c).

Table 5. Location of the 15 Descriptors around the Center of Origin of the 3D Loading Plots of the PCA Diagrams in Figure 1b (DSL) and d (FKL)

octant	DSL	FKL
1	fAromRings, fUnsatBonds	ALogP, HevAtoms, MW
2	ALogP, HevAtoms, MW	fAromRings, fUnsatBonds
3	fRotBonds	fHBD
4	fHBA, fHBD, fHetAtoms, PSA	fHBA, fHetAtoms, fRotBonds, PSA
5	fRings, fRingSys	–
6	–	fRings, fRingSys
7	fSP3C	FCFP4density, fHetRings
8	FCFP4density, fHetRings	fSP3C

followed-up hits of similar size (MW = 332−342 Da, Figure 4b).

PAINS Evaluation. The presence of nonspecific frequent hitters within screening libraries is a common problem associated with false positives from screening campaigns.[11] To investigate if any problematic compounds were present in our screening libraries, we followed the definitions of Baell and Holloway[9] which stated that screening compounds might be displaying PAINS behavior if reported active in more than 50% of the number of assays screened. Compounds within each screening library were grouped according to the number of assays in which each individual compound was reported as active (Table 6). Primary hits and followed-up hits were tabulated separately. Screened against seven different targets (Table 2), DSL had no individual compound reported as primary hits in more than three assays. At the level of followed-up hits, no compound was active in more than two assays (Table 6). Similarly, FKL was screened against ten different targets (Table 3), and no compound was active in more than five assays at the level of primary hits or followed-up hits (Table 6). These observations suggested that both DSL and FKL did not contain any compounds displaying PAINS behavior according to the definitions of Baell and Holloway.

In addition, we evaluated whether compounds containing structural motifs mapping to literature PAINS filters[9] were frequently reported as active in multiple assays using our screening data. We applied the PAINS substructure filters published by Baell and Holloway[9] to flag any compounds within these libraries that contained structural motifs which are likely to display PAINS behavior (Table 7 and Supporting Information). For DSL, 1725 compounds (2.9%) matching 97 literature PAINS structural motifs were flagged by the substructure filters as potential PAINS, whereas 50 compounds (1.5%) matching 9 literature PAINS structural motifs were flagged for FKL (Supporting Information Tables S1 and S2). Only 85 of the flagged 1725 compounds in DSL were reported as a primary hit, with 28 compounds also satisfying our followed-up hit criteria (Table 7). This illustrated that over 95% of the flagged compounds were inactive against all seven targets screened. Switching to FKL, 31 of the 50 flagged compounds were not active against any of the ten targets screened, which represented a 62% clean rate of these flagged compounds. Most of the remaining flagged compounds were only active in one or two assays.

Further, we assessed PAINS behavior on a structural motif level instead of on an individual compound level. For this analysis, we grouped the flagged compounds within each library according to the PAINS structural motifs and investigated in how many different assays representatives of each motif appeared as actives. Out of the 97 motifs present in DSL compounds, 55 motifs were considered underrepresented with fewer than five examples and were excluded from the following analysis (Supporting Information Table S1). No active compounds were reported for 19 of the remaining 42 motifs, while another 12 motifs contained compounds that were active only in one assay. Only one motif (5-membered alkylidene heterocycles, ene_five_het_B in Baell and Holloway)[9] contained compounds that were altogether reported as primary hits in more than half of the assays (Tables 8 and S1). In FKL, only two of the nine motifs present were reasonably represented by at least five examples, and none of these contained compounds that were altogether reported as primary hits in more than half of the assays (Tables 8 and S2). Since there were only a small number of flagged compounds that were classified as followed-up hits, we decided that the analysis of followed-up hits grouped into PAINS structural motifs would be inconclusive.

DISCUSSION

The efficiency of hit identification in automated screening relies heavily on the quality of the screening libraries used. There are numerous ways to evaluate the quality of a screening library. Here, we were interested in the utilization of chemical space represented by a diverse (DSL) and a kinase-focused (FKL) lead-like screening library and the distribution of screening hits within their respective chemical space. We also assessed whether any library compounds were displaying PAINS behavior according to the definitions of Baell and Holloway.[9]

Both libraries delivered screening hits across a range of targets (Tables 2 and 3). DSL had hit rates ranging from 0.005 to 1.21%, whereas the hit rates for FKL range from 0.03 to 12.9%, with the highest hit rates against protein kinases for

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

(a)

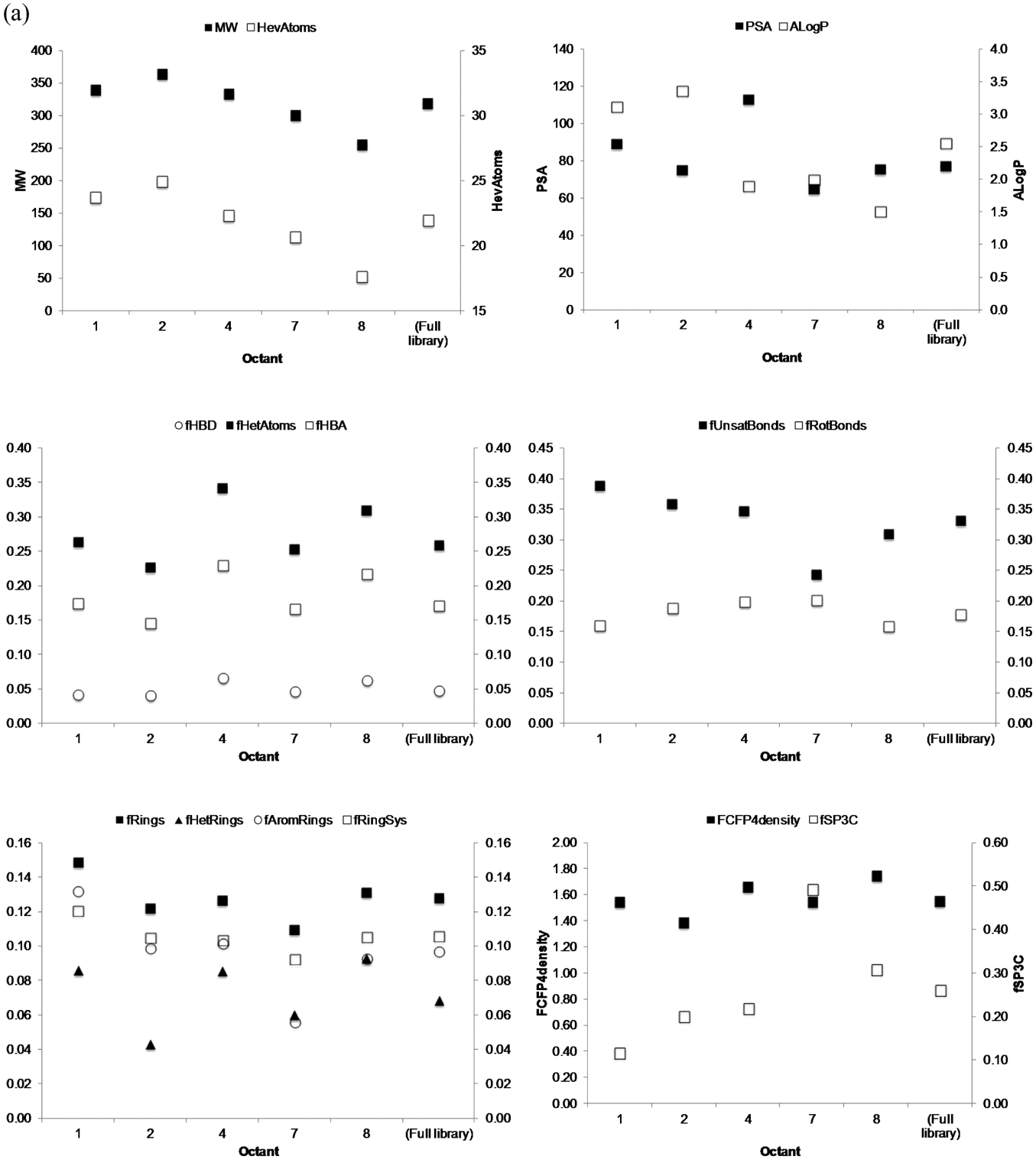


Figure 3. continued

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

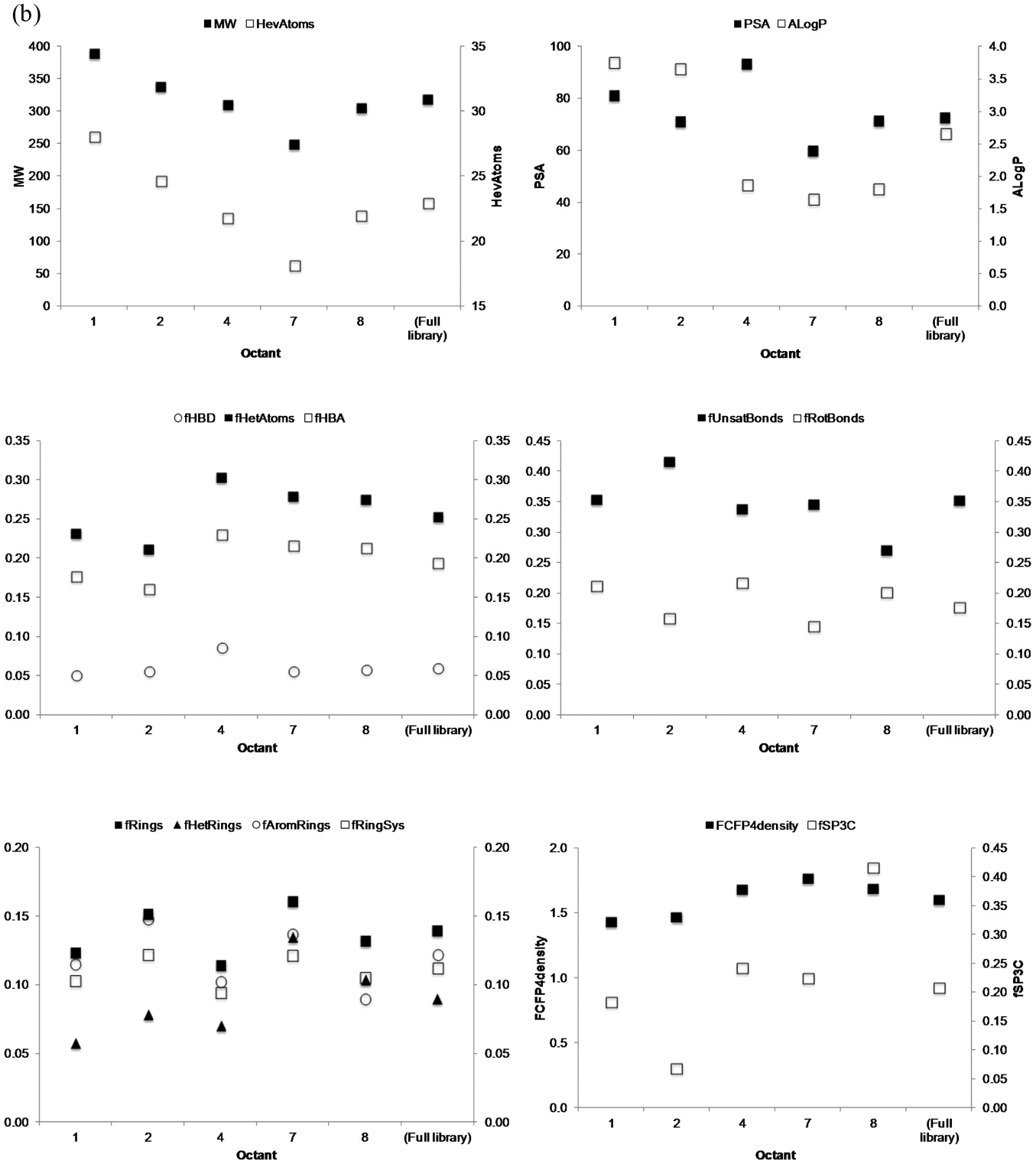


Figure 3. Plots showing the average values of each descriptor within octants 1, 2, 4, 7, 8, and that for the full library for DSL (a) and FKL (b).

which the library was originally designed.[13] These hit rates are comparable to those typically reported for screening campaigns,[35−37] especially considering that most of the targets apart from the protein kinases have not previously been subjected to automated screening. This indicates that the investigated libraries are overall suitable for hit discovery.

According to the chemical space analyses, both DSL and FKL libraries were able to deliver hits across the entire chemical space represented, and there were no apparent regions of chemical space where no hits could be found (Figure 1). This illustrates that the entire chemical space covered by these lead-like libraries can be utilized to probe interactions between proteins and small-molecule ligands. However, despite that hits were identified across the entire chemical space of the respective libraries, the distribution of hits was uneven when we analyzed the occupancy of each octant of the 3D-PCA plots

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

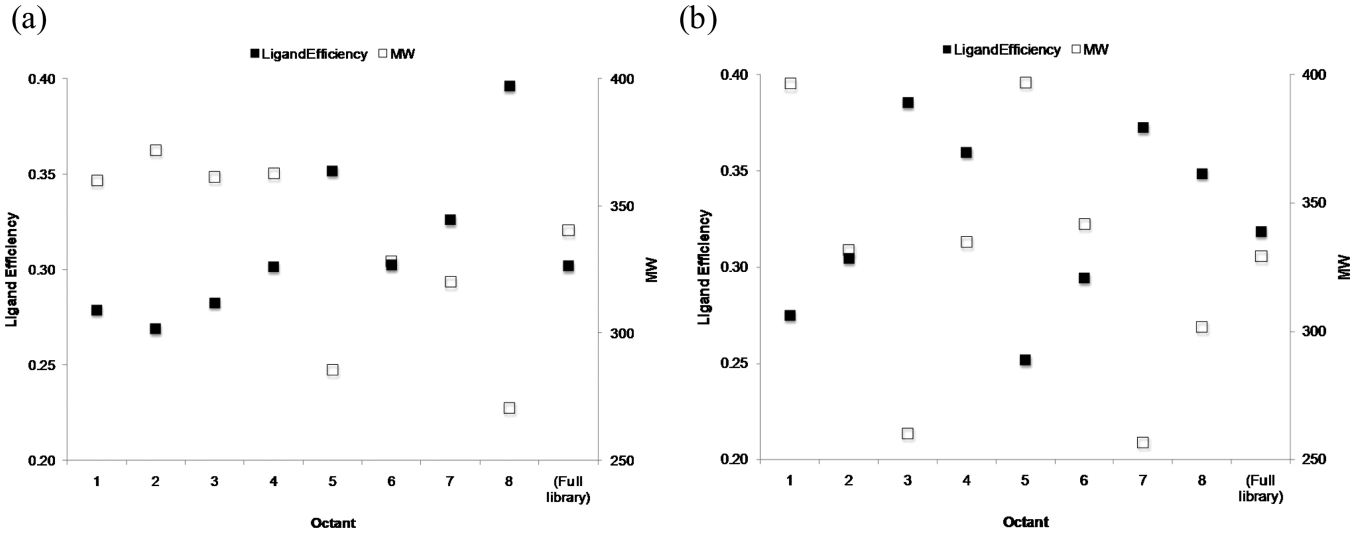


Figure 4. Plots showing the average ligand efficiency and molecular weight (MW) of followed-up hits within each octant and that for the full library for DSL (a) and FKL (b).

Table 6. Breakdown of the Number of Assays in Which Each Compound Was Reported Active As a Primary Hit or Followed-up Hit

	no. of assays							
DSL	0	1	2	3	4	5	6+	total
primary hits	57723	1657	58	5	0	0	0	59443
followed-up hits	59141	296	6	0	0	0	0	59443
FKL	0	1	2	3	4	5	6+	total
primary hits	2540	400	221	100	20	6	0	3287
followed-up hits	3032	223	27	5	0	0	0	3287

Table 7. Breakdown of the Number of Assays in Which Each Compound Flagged As PAINS Was Reported Active As a Primary Hit or Followed-up Hit

	no. of assays							
DSL	0	1	2	3	4	5	6+	total
primary hits	1640	76	9	0	0	0	0	1725
followed-up hits	1697	27	1	0	0	0	0	1725
FKL	0	1	2	3	4	5	6+	total
primary hits	31	10	6	2	1	0	0	50
followed-up hits	45	4	1	0	0	0	0	50

Table 8. Breakdown of the Number of Assays in Which Each PAINS Structural Motif Contained Compounds Reported As a Primary Hit

	no. of assays							
DSL	0	1	2	3	4	5	6+	total
all motifs	70	15	4	7	1	0	0	97
motifs with at least five representatives	19	12	4	6	1	0	0	42
FKL	0	1	2	3	4	5	6+	total
all motifs	2	0	5	1	0	1	0	9
motifs with at least five representatives	0	0	1	0	0	1	0	2

of each library individually (Figure 2). We observed, for both libraries, enrichments in the percentage of reported hits in octants occupied by heavy, lipophilic compounds, whereas the percentage of reported hits decreased in octants characterized by polar compounds or compounds containing a high fraction of sp^3-carbon atoms (Figure 3). Nonetheless, it should be emphasized that all of the hits are well within lead-like chemical space.

We propose that the observed uneven distribution of screening hits across the analyzed chemical space may be explained by the different intrinsic binding capabilities of compounds in the relevant octants. Hit compounds in the enriched octants are relatively lipophilic and bulky and contain a large fraction of aromatic rings (Figure 3). Accordingly, these molecules are rich in unsaturation and represent relatively flat molecular shapes. Owing to the relatively simple and generic molecular shapes, these compounds are more likely to participate in protein−ligand interactions without requiring a stringent spatial complement, therefore leading to higher hit rates. On the contrary, compounds with a high fraction of sp^3-carbon atoms represent more complex molecular shapes that require a higher shape complementarity at the protein−ligand interface to accommodate ligand binding.[33] Similarly, a complementary electrostatics match would be required for the successful binding of polar and heteroatom rich compounds. Hence, the hit rate obtained from polar compounds or compounds containing a high fraction of sp^3-carbon atoms would inevitably be lower than that from lipophilic aromatic compounds.

Recently, it was argued that compounds that are polar, heteroatom rich, or contain a high fraction of sp^3-carbon atoms represent better prospects for drug discovery, as candidates derived from these compounds are more likely to be successful in clinical trials.[33,38,39] Our analysis demonstrated that these compounds are also better lead candidates in terms of average ligand efficiency, exceeding on average the 0.30 kcal mol^{-1} per heavy atom cutoff that is generally considered favorable for developing a potent, Rule-of-Five compliant drug candidate (octants 4, 7, and 8; Figure 4).[34] It would therefore be desirable to increase the number of hits obtained from these octants. In order to attain this, we suggest that the entire screening library should not be evenly distributed across the octants but instead be enriched in compounds from the underrepresented octants

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

to achieve a more even distribution of screening hits across the entire chemical space represented.

In addition to library composition, we also envisage that a departure from screening at the same fixed molar concentration for all library compounds in a single screening campaign may help balance the distribution of screening hits from bias toward heavy, lipophilic compounds that on average have comparably lower ligand efficiency (octants 1 and 2, Figure 4). Since smaller compounds tend to display a lower potency, the commonly used screening paradigm of one-concentration-fits-all favors the identification of heavy compounds, whereas smaller compounds are disadvantaged even when all compounds are within lead-like chemical space.[40] If compounds are screened at variable concentrations, with higher concentrations used for smaller compounds to match with their theoretical binding capacity,[41] screening hits with lower potency but higher ligand efficiency would no longer be discriminated.

Both libraries are free from compounds displaying PAINS behavior according to the definitions of Baell and Holloway (Tables 6 and 7).[9] As frequent hitters are a common source of false positives in screening campaigns,[10] this is a surprisingly positive result. It is noteworthy that in the classification criteria used for primary hits and followed-up hits, compounds which might be interfering with a certain assay readout technology (for example compounds that absorb light at a certain wavelength of a colorimetric assay, or compounds which displayed quenching behavior in a fluorescence assay) were excluded as primary hits from the corresponding assays in the first place. Hence, the presented analysis of PAINS should be clean from assay-dependent problematic compounds. When compiling the libraries, apart from removing obviously colored compounds, no specific filters were used to remove potentially promiscuous compounds.[13] However, reactive compounds that potentially bear toxicity issues were discarded. As there is some overlap between these filters and the PAINS motifs, it appears that this also helped to improve the libraries in terms of promiscuous behavior.

Even though some compounds were still flagged to contain PAINS structural motifs, upon detailed analysis, the majority of these compounds did not show any activity against the panel of targets screened (Table 7 and the Supporting Information). This was also valid when the analysis was carried out on a structural motif level instead of on an individual compound level (Table 8 and the Supporting Information). Thus, in our hands, many of the reasonably represented PAINS structural motifs in our libraries appeared to be less of a nuisance in biochemical screens for enzyme assays than suggested previously by others.[9,42] For the purpose of enhancing the diversity of a screening library, we therefore consider it justifiable to include compounds containing PAINS structural motifs that were demonstrated to be relatively clean in our analysis, in particular when such compounds contain additional scaffolds that are otherwise not commercially available without the PAINS substituents. However, such compounds should be annotated in the library to ensure that the absence of promiscuous behavior is rigorously verified prior to any optimization efforts.

CONCLUSIONS

Using screening data from two lead-like screening libraries against 15 enzyme targets, we demonstrated that both libraries delivered hits across a range of targets. The screening hits spanned the entire lead-like chemical space covered by these libraries, although the distribution of screening hits was found to be uneven. With observed enrichments of screening hits that are at the higher end of the molecular weight and lipophilicity spectrum for lead-like compounds, we propose that screening libraries should in the future be enriched in polar, aliphatic compounds. In conjunction with the introduction of variable concentrations screening, we envisage that these could rectify the uneven distribution of hits observed. Such a movement in future screening library design should assist in discovering a higher proportion of screening hits with higher ligand efficiency and properties that have recently been suggested to lead to better selectivity and reduced likelihood of promiscuity, thereby maximizing potential success in clinical trials.

In addition, our analysis suggests a less stringent approach in the application of the literature PAINS filters in removing screening compounds. Both screening libraries were shown to be clean from any PAINS behavior according to the literature definitions. Even though some compounds were flagged as PAINS, the analysis on reasonably represented structural motifs demonstrated that some of these motifs appeared to be less problematic than previously suggested. Although compounds flagged by these PAINS structural motifs may not represent the top candidates for optimization into a drug when there are a large number of screening hits available, it is arguable whether such compounds should be completely excluded from a screening library. This is particularly relevant in diverse screening libraries that are compiled for screening against a wide spectrum of targets and phenotypes, since challenging screening campaigns might not always achieve high hit rates. We therefore consider it justifiable to retain compounds containing PAINS motifs demonstrated to be apparently clean in this study to maximize the chemical diversity in a screening library.

EXPERIMENTAL PROCEDURES

Descriptor Calculations. The 15 descriptors were calculated using Pipeline Pilot professional client 8.0 (Accelrys, Inc.) applying the definitions in the software unless stated otherwise. All categorical descriptors with discrete unit values were normalized relative to the number of heavy atoms unless stated otherwise.

A heteroatom was defined as the elements S, O, or N. An unsaturated bond was defined as a bond with a bond order greater than one. A heterocycle was defined as a ring containing S, O, or N in the fragment that resulted from generating fragments by rings. An sp^3-hybridized carbon atom was defined as any carbon atom which has an atom hybridization of sp^3 according to Pipeline Pilot calculations. The fraction of sp^3-hybridized carbon atoms was normalized relative to the total number of carbon atoms in the same molecule.[33] FCFP4density was defined as the ratio between the number of bits in the FCFP4 fingerprint generated and the number of heavy atoms.[32] Ligand efficiency of followed-up hits was determined using the IC_{50} value (the most potent IC_{50} was chosen for calculations when a compound has IC_{50} values for more than one target) following the equation

$$\text{ligand efficiency} = -RT\ \ln(IC_{50})/\text{number of heavy atoms}$$

where $R = 1.98 \times 10^{-3}$ kcal K^{-1} mol^{-1} and $T = 300$ K.[34] (IC_{50} values were typically determined with a substrate concentration close to K_m so that $IC_{50} \approx K_i$ assuming competitive inhibition.)

Chemical Space Analysis. The 3D-PCA plots were generated using Simca-P+ 12.0.1 (Umetrics). The descriptor

dx.doi.org/10.1021/ci300382f | J. Chem. Inf. Model. 2013, 53, 534−544

matrix was normalized to unit variance before carrying out PCA using the PCA-X option under standard settings. The number of principal components was based on automatic cross-validation within the software.

PAINS Analysis. The literature PAINS filters in SLN format (Tables S6, S7, and S9 in the Supporting Information from the work of Baell and Holloway)[9] were applied using Sybyl-X 1.2 (Tripos). The flagged compounds were mapped to individual PAINS substructure motifs using in-house Python scripts.

■ ASSOCIATED CONTENT

Supporting Information

Figures of the distribution of primary and followed-up hits among enzyme targets screened within each octant and tables listing the PAINS structural motifs that were present in DSL and FKL. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: brenk@uni-mainz.de. Phone: +49 (6131) 39-25727.

Present Addresses

[‡]N.Y.M.: Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, 15 Cotswold Road, Belmont, London SM2 5NG, U.K.

[§]R.B.: Johannes Gutenberg-Universität Mainz, Institut für Pharmazie und Biochemie, Staudinger Weg 5, 55128 Mainz, Germany.

Author Contributions

[†]N.Y.M. and S.M. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust (WT083481 and WT092340). We thank the Scottish Universities Life Sciences Alliance (SULSA) for financial support of Pipeline Pilot software (Grant HR07019), Daniel Muthas for advice on the PCA, and David Blair, Iain Collie, Manu De Rycker, Ian Gilbert, David Gray, Raffaella Grimaldi, Irene Hallyburton, Daniel James, Dhananjay Joshi, Stuart McElroy, Naomi Tidten-Luksch, and Leah Torrie for screening data collection.

■ REFERENCES

(1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; 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 Discov.* **2011**, *10*, 188−195.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(3) 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.

(4) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.

(5) Miller, J. L. Recent developments in focused library design: Targeting gene-families. *Curr. Top. Med. Chem.* **2006**, *6*, 19−29.

(6) Nadin, A.; Hattotuwagama, C.; Churcher, I. Lead-oriented synthesis: a new opportunity for synthetic chemistry. *Angew. Chem.* **2012**, *51*, 1114−1122.

(7) Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.

(8) Murray, C. W.; Verdonk, M. L.; Rees, D. C. Experiences in fragment-based drug discovery. *Trends Pharmacol. Sci.* **2012**, *33*, 224−232.

(9) 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.

(10) Thorne, N.; Auld, D. S.; Inglese, J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr. Opin. Chem. Biol.* **2010**, *14*, 315−324.

(11) Sink, R.; Gobec, S.; Pecar, S.; Zega, A. False positives in the early stages of drug discovery. *Curr. Med. Chem.* **2010**, *17*, 4231−4255.

(12) Che, J.; King, F. J.; Zhou, B.; Zhou, Y. Chemical and biological properties of frequent screening hits. *J. Chem. Inf. Model.* **2012**, *52*, 913−926.

(13) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *Chemmedchem* **2008**, *3*, 435−444.

(14) Mok, N. Y.; Brenk, R. Mining the ChEMBL database: An efficient chemoinformatics workflow for assembling an ion channel-focused screening library. *J. Chem. Inf. Model.* **2011**, *51*, 2449−2454.

(15) Frearson, J. A.; Brand, S.; McElroy, S. P.; Cleghorn, L. A.; Smid, O.; Stojanovski, L.; Price, H. P.; Guther, M. L.; Torrie, L. S.; Robinson, D. A.; Hallyburton, I.; Mpamhanga, C. P.; Brannigan, J. A.; Wilkinson, A. J.; Hodgkinson, M.; Hui, R.; Qiu, W.; Raimi, O. G.; van Aalten, D. M.; Brenk, R.; Gilbert, I. H.; Read, K. D.; Fairlamb, A. H.; Ferguson, M. A.; Smith, D. F.; Wyatt, P. G. N-myristoyltransferase inhibitors as new leads to treat sleeping sickness. *Nature* **2010**, *464*, 728−732.

(16) Cleghorn, L. A.; Woodland, A.; Collie, I. T.; Torrie, L. S.; Norcross, N.; Luksch, T.; Mpamhanga, C.; Walker, R. G.; Mottram, J. C.; Brenk, R.; Frearson, J. A.; Gilbert, I. H.; Wyatt, P. G. Identification of inhibitors of the Leishmania cdc2-related protein kinase CRK3. *Chemmedchem* **2011**, *6*, 2214−2224.

(17) Patterson, S.; Alphey, M. S.; Jones, D. C.; Shanks, E. J.; Street, I. P.; Frearson, J. A.; Wyatt, P. G.; Gilbert, I. H.; Fairlamb, A. H. Dihydroquinazolines as a novel class of Trypanosoma brucei trypanothione reductase inhibitors: discovery, synthesis, and characterization of their binding mode by protein crystallography. *J. Med. Chem.* **2011**, *54*, 6514−6530.

(18) Brand, S.; Cleghorn, L. A.; McElroy, S. P.; Robinson, D. A.; Smith, V. C.; Hallyburton, I.; Harrison, J. R.; Norcross, N. R.; Spinks, D.; Bayliss, T.; Norval, S.; Stojanovski, L.; Torrie, L. S.; Frearson, J. A.; Brenk, R.; Fairlamb, A. H.; Ferguson, M. A.; Read, K. D.; Wyatt, P. G.; Gilbert, I. H. Discovery of a novel class of orally active trypanocidal N-myristoyltransferase inhibitors. *J. Med. Chem.* **2012**, *55*, 140−152.

(19) Tidten-Luksch, N.; Grimaldi, R.; Torrie, L. S.; Frearson, J. A.; Hunter, W. N.; Brenk, R. IspE inhibitors identified by a combination of in silico and in vitro high-throughput screening. *PloS one* **2012**, *7*, e35792.

(20) Malcolm, B. A. The Picornaviral 3C proteinases - cysteine nucleophiles in serine proteinase folds. *Protein Sci.* **1995**, *4*, 1439−1445.

(21) Devaney, E.; O'Neill, K.; Harnett, W.; Whitesell, L.; Kinnaird, J. H. Hsp90 is essential in the filarial nematode Brugia pahangi. *Int. J. Parasitol.* **2005**, *35*, 627−636.

(22) Hammarton, T. C.; Kramer, S.; Tetley, L.; Boshart, M.; Mottram, J. C. Trypanosoma brucei Polo-like kinase is essential for basal body duplication, kDNA segregation and cytokinesis. *Mol. Microbiol.* **2007**, *65*, 1229−1248.

(23) Ojo, K. K.; Gillespie, J. R.; Riechers, A. J.; Napuli, A. J.; Verlinde, C. L. M. J.; Buckner, F. S.; Gelb, M. H.; Domostoj, M. M.; Wells, S. J.; Scheer, A.; Wells, T. N. C.; Van Voorhis, W. C. Glycogen synthase kinase 3 is a potential drug target for African trypanosomiasis therapy. *Antimicrob. Agents Chemother.* **2008**, *52*, 3710−3717.
(24) Stokes, M. J.; Guther, M. L. S.; Turnock, D. C.; Prescott, A. R.; Martin, K. L.; Alphey, M. S.; Ferguson, M. A. J. The synthesis of UDP-N-acetylglucosamine is essential for bloodstream form Trypanosoma brucei in vitro and in vivo and UDP-N-acetylglucosamine starvation reveals a hierarchy in parasite protein glycosylation. *J. Biol. Chem.* **2008**, *283*, 16147−16161.
(25) Dvorin, J. D.; Martyn, D. C.; Patel, S. D.; Grimley, J. S.; Collins, C. R.; Hopp, C. S.; Bright, A. T.; Westenberger, S.; Winzeler, E.; Blackman, M. J.; Baker, D. A.; Wandless, T. J.; Duraisingh, M. T. A plant-like kinase in Plasmodium falciparum regulates parasite egress from erythrocytes. *Science* **2010**, *328*, 910−912.
(26) Ma, J. T.; Benz, C.; Grimaldi, R.; Stockdale, C.; Wyatt, P.; Frearson, J.; Hammarton, T. C. Nuclear DBF-2-related kinases are essential regulators of cytokinesis in bloodstream stage Trypanosoma brucei. *J. Biol. Chem.* **2010**, *285*, 15356−15368.
(27) Lazarus, M. B.; Nam, Y. S.; Jiang, J. Y.; Sliz, P.; Walker, S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **2011**, *469*, 564−567.
(28) de Alencar, N. A. N.; Sousa, P. R. M.; Silva, J. R. A.; Lameira, J.; Alves, C. N.; Marti, S.; Moliner, V. Computational analysis of human OGA structure in complex with PUGNAc and NAG-thiazoline derivatives. *J. Chem. Inf. Model.* **2012**, *52*, 2775−2783.
(29) Torrie, L. S.; Wyllie, S.; Spinks, D.; Oza, S. L.; Thompson, S.; Harrison, J. R.; Gilbert, I. H.; Wyatt, P. G.; Fairlamb, A. H.; Frearson, J. A. Chemical validation of Trypanothione synthetase A potential drug target for human Trypanosomiasis. *J. Biol. Chem.* **2009**, *284*, 36137−36145.
(30) Zhang, J. H.; Chung, T. D. Y.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **1999**, *4*, 67−73.
(31) Shoichet, B. K. Interpreting steep dose-response curves in early inhibitor discovery. *J. Med. Chem.* **2006**, *49*, 7274−7277.
(32) Schuffenhauer, A.; Brown, N.; Selzer, P.; Ertl, P.; Jacoby, E. Relationships between molecular complexity, biological activity, and structural diversity. *J. Chem. Inf. Model.* **2006**, *46*, 525−535.
(33) 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.
(34) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430−431.
(35) Posy, S. L.; Hermsmeier, M. A.; Vaccaro, W.; Ott, K. H.; Todderud, G.; Lippy, J. S.; Trainor, G. L.; Loughney, D. A.; Johnson, S. R. Trends in kinase selectivity insights for target class-focused library screening. *J. Med. Chem.* **2011**, *54*, 54−66.
(36) Xi, H. L.; Lunney, E. A. The design, annotation, and application of a kinase-targeted library. *Methods Mol. Biol.* **2011**, *685*, 279−291.
(37) Posner, B. A.; Xi, H. L.; Mills, J. E. J. Enhanced HTS hit selection via a local hit rate analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2202−2210.
(38) Leeson, P. D.; St-Gallay, S. A. The influence of the 'organizational factor' on compound quality in drug discovery. *Nat. Rev. Drug Discovery* **2011**, *10*, 749−765.
(39) Dandapani, S.; Marcaurelle, L. A. Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.* **2010**, *6*, 861−863.
(40) Nowlin, D.; Bingham, P.; Berridge, A.; Gribbon, P.; Laflin, P.; Sewing, A. Analysing the output from primary screening. *Comb. Chem. High Throughput Screening* **2006**, *9*, 331−337.
(41) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9997−10002.
(42) Huth, J. R.; Mendoza, R.; Olejniczak, E. T.; Johnson, R. W.; Cothron, D. A.; Liu, Y. Y.; Lerner, C. G.; Chen, J.; Hajduk, P. J.; ALARM, N. M. R. A rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* **2005**, *127*, 217−224.