

Broad Coverage of Commercially Available Lead-like Screening Space with Fewer than 350,000 Compounds

Jonathan B. Baell^{*,†,‡,§}

[†]Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University (Parkville Campus), 381 Royal Parade, Parkville, VIC 3052, Australia

[‡]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

[§]Department of Medical Biology, University of Melbourne, Parkville, Victoria 3010, Australia

Supporting Information

ABSTRACT: In establishing what we propose is the globe's highest quality collection of available screening compounds, it is convincingly shown that the globe's pool of such compounds is extremely shallow and can be represented by fewer than 350,000 compounds. To support our argument, we discuss and fully disclose our extensive battery of functional group filters. We discuss the use of PAINS filters and also show the effect of similarity exclusion on structure–activity relationships. We show why limited analogue representation requires screening at higher concentrations to capture hit classes for difficult targets that otherwise may be prosecuted unsuccessfully. We construct our arguments in a structurally focused manner to be most useful to medicinal chemists, the key players in drug discovery.



■ INTRODUCTION

One of the best ways to find new starting points for drug discovery involves high throughput screening.¹ Traditionally the domain of big pharma, this technology has become established in academic settings over the past decade.² Whereas pharmaceutical companies usually have access to large proprietary screening libraries, academic groups have generally been reliant on establishing their HTS libraries from commercial sources. This demand has been readily met by chemical vendors who were already working to supply compounds for industry partners. Vendors usually have sophisticated file processing capabilities and, presumably in response to demand, confine the majority of their compounds to within Lipinski³ criteria.⁴ When we established our Stage 1 library in 2003, we assumed that commercially available libraries would additionally be largely devoid of functional groups flagged as problematic for drug development.^{5–7} We were fortunate to have the facility to undertake sophisticated file processing, and in establishing our library from a selection of vendors we were initially surprised that significant numbers of compounds were recognized by and removed with the functional group filters that we had defined. We also observed many compound classes that were excessively populated with very large numbers (up to 300 for example) of very similar analogues.

Over recent years we have formed the view that it is entirely acceptable that chemical vendors sell compounds that may fail Lipinski and certain functional group screening criteria because buyers may want compounds for diverse uses where such

criteria may be less relevant, for example anti-infectives, protein–protein interaction inhibitors, tool compounds or indeed uses less related to medicinal chemistry. Indeed, the NIH Molecular Libraries Initiative was established using relatively lenient criteria for these sorts of reasons.⁸ Furthermore, clarity is often lacking as to why a given functional group may be viewed unfavorably: is it because of association with false positives or perhaps because of the potential for poor pharmacokinetic (PK) behavior or perhaps because of the potential for downstream toxicity? How limiting each of these characteristics is perceived to be may be quite subjective to the extent that some researchers might be comfortable purchasing particular compounds for screening that others would not. Hence we think it is appropriate that there be a broad representation of compounds in vendor libraries. It is also reasonable for vendors to sell large numbers of similar analogues because purchase of these represents a facile way to establish first generation SAR for a given screening hit: however, one would not necessarily want to purchase all such analogues in establishing an HTS library to the exclusion of alternative compound class representation.

For these reasons, there is considerable onus on academic groups wishing to establish new HTS libraries to either interact closely with vendors to select precisely the types of compounds they want or to develop batteries of filters and undertake extensive file processing themselves in order to select the most

Received: September 27, 2012

Published: November 30, 2012

appropriate compounds for their screening needs. For more than a decade, we have been increasing the sophistication of our compound filters over the course of six individual library expansion events. These events are shown in Table 1 and represent a total of some 405,000 compounds.

Table 1. Different HTS Libraries Established over Time and Selection Principles

library name (year established)	number of compounds	number of vendors	broad selection principles ^a	PAINS filter? ^b	similarity filter? ^c
Stage 1 (2003)	93,000	4	lead-like	N	Y
Stage 2 (2003)	30,000	1	diversity	N	N
Stage 3 (2007)	17,000	20	clustering	N	Y
Stage 4 (2007)	15,000	1	lead-like	Y	Y
Stage 5 (2007)	136,000	2	lead-like	Y	Y
Stage 6 (2010)	114,000	10	lead-like	Y	Y

^aLead-like criteria generally as follows: mw 150–400; rings 1–4; cLogP_{max} 5; rot. bonds_{max} 10; chiral_{max} 3; HBD_{max} 3–5; HBA_{max} 6–8; inappropriate functional groups; ^bFilters previously summarized elsewhere.⁹ Early versions were used for Stages 4 and 5. ^cGenerally analogues more than 85% similar (or 90% similar for a portion of the Stage 5 library) as defined by the Tanimoto coefficient. The Stage 2 and 3 libraries were sourced via independent 3rd parties and contain some compounds more than 90% similar to those in Stages 4, 5 and 6.

It can be seen that our guiding philosophy has been to select compounds based on lead-likeness and to exclude compounds more than 85–90% similar to any others already selected. As well as using similarity criteria, in establishing our most recent Stage 6 library during 2010, we developed a vastly expanded set of functional group filters to capture unattractive compounds that hitherto had escaped our filters defined in 2007. This Stage 6 library was also the first time we applied our refined PAINS filters⁹ to exclude nuisance compounds from purchase.

Embodied in our Stage 6 library is therefore a rich history and understanding of available screening chemistry. The purpose of this article is several-fold. First, to elaborate on the rationale of functional group filter use in establishing general purpose HTS libraries of as high quality as possible. Second, to expand commentary on the use of PAINS filters, which have elicited significant scientific debate.^{10–12} Third, to analyze the profound effect of Tanimoto similarity removal in a manner more useful to medicinal chemists than has been done hitherto. Fourthly and most fundamentally, to disclose the revelation that our criteria in the first instance led to exhaustion of commercially available compounds from vendors representing 80% of available lead-like space before exhaustion of our budget, with fewer than 100,000 compounds remaining in the selection pool. This latter observation will be that which is discussed first.

RESULTS

The high rate of attrition that results from our filtering processes is starkly revealed in the schematic of Figure 1 for a typical vendor with a catalogue of around 400,000 compounds. Here, application of our first lead-like filter resulted in the removal of around 150,000 compounds. Then, a functional

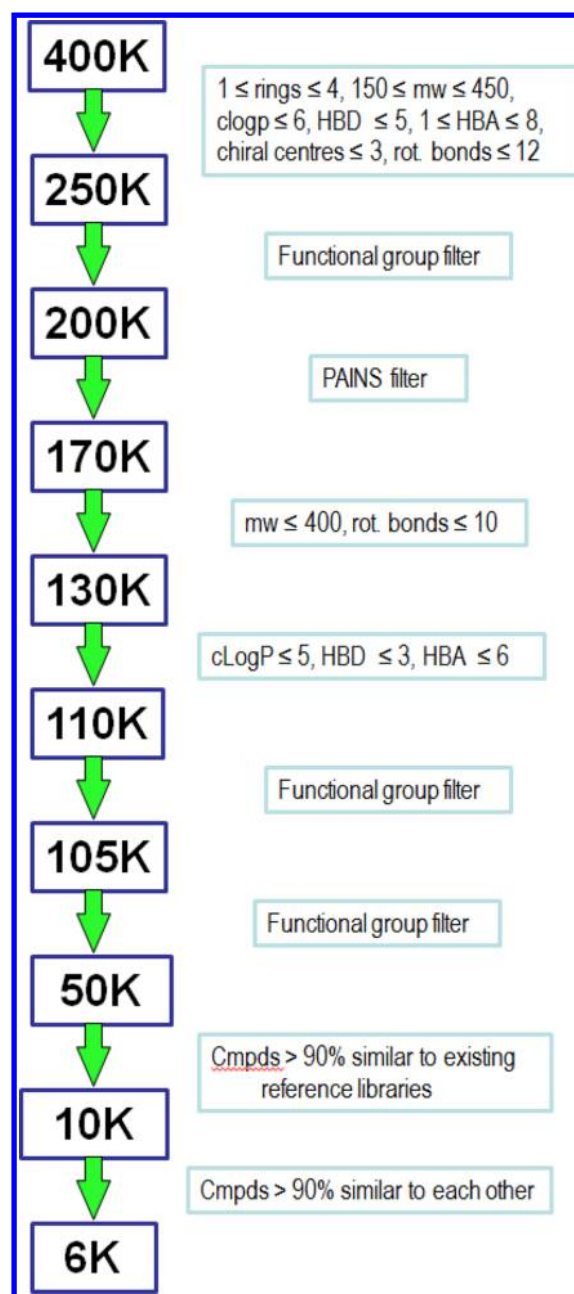


Figure 1. Schematic illustration of the attrition of a given vendor's library of 400,000 compounds with progressive filtering on lead-likeness and removing similar compounds.

group filter for nondrug-like groups was applied, followed by our PAINS filters,⁹ leaving 170,000 compounds. Progressive filtering on harsher lead-like criteria removed another 60,000 compounds, leaving 110,000 compounds. At this point, intensive scrutiny of remaining compounds led to the development of significant numbers of new functional group filters that had not previously been applied to any of our earlier libraries.

This left us with 50,000 compounds. Then, all compounds more than 90% similar to any in our more recent (2007) libraries were removed, leaving 10,000 compounds. Finally, all compounds more than 90% similar to each other within this remaining set of 10,000 compounds were removed, leaving us with 6,000 compounds. Hence, for this particular vendor, a remarkable attrition rate of 98.5% was observed.

Table 2. Triage of Each Vendor Library, Showing Number of Compounds Remaining after Operation for the Ten Vendors A–J

entry	vendor									
	A	B	C	D	E	F	G	H	I	J
1 ^a	1,168,625	619,514	1,165,361	354,629	449,998	239,674	193,379	289,552	392,499	61,623
2 ^b	871,476	493,927	716,111	253,876	298,131	162,524	135,282	215,617	249,722	50,267
3 ^c	712,017	406,377	636,166	207,573	211,023	121,975	94,261	187,026	195,628	36,779
4 ^d	689,082	385,050	606,905	194,152	188,128	111,348	86,524	177,707	172,967	35,094
5 ^e	512,297	304,827	374,374	147,582	149,586	91,121	70,776	120,135	131,556	31,832
6 ^f	475,099	272,458	320,820	132,832	126,242	78,254	61,661	107,840	111,754	28,122
7 ^g	452,909	262,116	304,660	125,723	113,638	71,595	55,317	105,572	104,268	25,150
8 ^h	262,215	161,726	179,290	67,522	64,141	40,525	30,117	60,592	47,901	14,233

^aAll sd files of available compounds combined, translated into sybyl line notation, stripped of salts and neutralized. ^bMixtures, metals, isotopes removed, $1 \leq \text{rings} \leq 4$, $150 \leq \text{mw} \leq 450$, $\text{clogp} \leq 6$, $\text{HBD} \leq 5$, $1 \leq \text{HBA} \leq 8$, chiral centers ≤ 3 , nonterminal rotatable bonds ≤ 12 . ^cResult after filtering free of groups in Filter 1 (SI, pp 55–59). ^dResults after filtering free of refined PAINS filters as per our publication.⁹ This essentially removed 162K compounds. ^e $\text{mw} \leq 400$, nonterminal rotatable bonds ≤ 10 . ^f $\text{cLogP} \leq 5$, $\text{HBD} \leq 3$, $\text{HBA} \leq 6$. ^gResult after filtering free of groups in Filter 2 (SI, pp 60–61). ^hResult after filtering free of groups in Filter 3 (SI, pp 62–80).

Table 3. Operations Involving Similarity Filtering: Selection of All Compounds That Are Not Excessively Analogous to Any Others Processed Prior

entry		number of compounds									
1	base set ^a	262,215	161,726	179,290	67,522	64,141	40,525	30,117	60,592	47,901	14,233
2	reference database size ^b	180,759	301,412	335,335	344,540	353,605	359,179	364,390	367,766	377,515	383,099
3	Tan90 set ^c	212,395	49,090	23,237	15,621	7,176	6,242	3,956	21,764	9,447	8,046
4	Tan90 set ^d	120,653	33,943	9,205	9,065	5,574	5,211	3,376	9,749	5,584	6,587

^aThis corresponds to entry 8 in Table 2. ^bInitial reference database of 180,759 comprises the sum of our Stage 2, 3, 4 and Stage 5 libraries. The Stage 2 and 3 libraries were provided by 3rd parties and so compounds therein abide by their selection criteria; from these two libraries we excluded about 18K compounds from further consideration and these are absent from the initial reference database. The next reference database of 301,412 comprises the sum of the initial reference database of 180,759 compounds and vendor A compounds remaining that are first less than 90% similar to this (entry 3 with 212,395 compounds) and then also to each other (entry 4 with 120,653 compounds). The subsequent reference databases increase in size analogously after processing each preceding vendor. ^cCompounds remaining from the number for each vendor's base set (entry 1) after removal of all compounds more than 90% similar to those in the reference database. ^dCompounds remaining from the number for each vendor shown in entry 3 after removal of all compounds more than 90% similar to each other.

This process was repeated for several other vendors, with similar results. At this point, we had far fewer compounds available than we had budgeted for, and so we continued to process new vendor catalogues until we eventually discovered a vendor – the tenth – that appeared to occupy significantly new chemical diversity space, being able to supply a relatively large number of compounds. In order to maximize our buying power, we repeated the entire process, but starting with this particular vendor in order to be able to select the maximum number of compounds. The subsequent vendor order chosen for processing was dictated by similar criteria. The details for this process up to the point of similarity filtering are listed in Table 2, while the details for the subsequent operations involving the similarity filtering are shown in Table 3. We have described this process in the way it was undertaken by us to lead to our final selection of compounds. Of course, all functional group filters could in principle be combined in a single filtering step.

At this point, all selected vendor compounds (entry 4 in Table 3) were combined and further processed. This is shown schematically in Figure 2, where the 208K combined compounds were reduced to 200K after removing those more than 99% similar to the 2003 Stage 1 Library. The reason for the lenient similarity filter was because the Stage 1 library was about to be exhausted, and it would be inappropriate to remove compounds based on similarity to compounds in this library. However, neither did we wish to screen duplicate compounds while this Stage 1 library was still “alive”, hence the Tanimoto coefficient cutoff of 0.99.

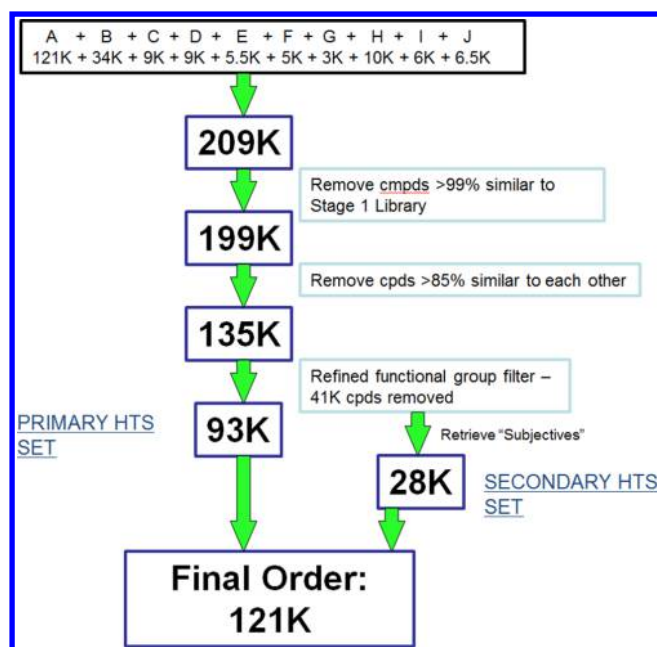


Figure 2. Schematic representation of the final processing of collated, remaining vendor compounds.

At this stage, we now removed all compounds that were more than 85% similar to each other, and this removed 65,000 compounds, leaving us with 135,000 compounds. Scrutiny of random exports of 1000 compounds revealed a significant

Table 4. Final Vendor Orders^a

vendor	available compounds in primary selection set (original number requested)	available compounds in secondary selection set (original number requested)	cost (\$ μ mol) relative to vendor A	resupply details
A	52,092 (52,707)	17,631 (17,753)	A	poor
B	15,643 (18,643)	2,448 (2,837)	1.3–1.9A	poor
C	3,943 (4,528)	1,014 (1,174)	2.5A	good
D	3,384 (3,980)	1,129 (1,242)	1.3A	poor
E	1,680 (2,127)	630 (766)	2.3A	inadequate
F	1,666 (1,706)	648 (660)	2.3A	good
G	1,563 (1,593)	487 (487)	2.5A	poor
H	3,883 (3,902)	1,704 (1,712)	1.3A	poor
I	2,453 (2,453)	828 (872)	2.8A	poor
J	1,666 (1,808)	575 (626)	2.6A	good
total available	87,973	27,094		

^aTotal number of compounds at end of processing = 121,543 (see Table S1 in the SI, p 91); total number requested (87,973 + 27,094) after 33 spurious compounds were retrieved = 121,576; total actually available at time of order = 115,067; final number of compounds received = 113,946. Vendors were represented by Asinex, ChemBridge, ChemDiv, Enamine, InterBioScreen, Life Chemicals, Maybridge, Specs, and TimTec. In order to provide some protection, these are listed alphabetically and not related to the order of vendors A–J used herein.

number of compounds that could be regarded as of secondary quality because of certain functional groups they contained. Filters were developed to identify these, and some 41,150 compounds were removed. Subsequently, it was determined that the budget was more than sufficient to purchase the remaining 93,415 refined compound set, and so 28,128 compounds were extracted from the rejected set of 41,150 compounds on the basis that, were they the only hit for a given screen, some medicinal chemists would be comfortable in progressing them. Such compounds include coumarins, hydrazones, thioethers, aminothiazoles, chromones, oximes, hydantoin, and sulfamides (see part ii of Filter 4 in the SI, pp 81–87). The detailed numbers for this process are shown in the SI (p 91).

Each vendor's contribution to this remaining compound set was ascertained and orders were placed. For some vendors such as B, C, and E there were quite a few compounds that had become unavailable by the time of purchase, this being 21% in the case of E (primary selection set), which was cause for concern, as this company also had a nonexistent resupply policy.

As shown in Table 4, a relatively good price was able to be negotiated when a large number of compounds could be selected from the one vendor. The cost of some vendor compounds was up to 2.8 times higher than the lowest priced vendor A. Also, there were only two vendors whose resupply philosophy we considered to be sound, this philosophy being that around 40 mg stock was set aside and accessible exclusively to prior buyers of that compound for the purposes of resupply of ca. 1–2 mg for secondary assay of a confirmed screening hit. *We believe this issue of resupply is in urgent need of attention by vendors and will provide a competitive edge to those vendors willing to better guarantee resupply.* Despite reservations on some resupply protocols, we purchased from all selected vendors, primarily because we had the budget to do so and saw great value in the quality of selected compounds remaining. In the majority of cases, vendors were able to supply compounds in our requested format as barcoded, ABgene microtubes in dry film.

Analysis of the new Stage 6 library revealed, as expected, favorable physicochemical properties, with the following averages: molecular weight 328, #rings 2.9, #chiral centers 0.3, #HBD 1.4, #HBA 3.3, #aromatic rings 2.3, cLogP 3.0, PSA

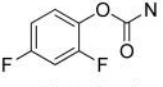
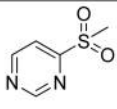

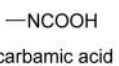




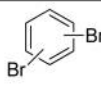
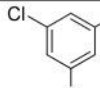
59 Å² (see the SI for histograms, pp 2–4). The relative degree of carbon saturation¹³ was lower than that typical of more optimized compounds but entirely acceptable in a screening hit and similar to other analyses of vendor-supplied compounds (average Fsp³ = 0.30).¹⁴ By virtue of our selection criteria, this library also contains a sizable, high quality Ro3 sublibrary¹⁵ of 14,222 compounds, along with reduced numbers of excessively flat and similar fragments. Functional group analysis (see the SI, p 90) indicated the following numbers: acylureas – cyclic such as hydantoin and pyrimidine diones (1311); acylureas – linear (727); alcohols (3766); amides (62808); amines (18241); anilines – primary, secondary, tertiary, and cyclic (6103); aromatic nitrogen (61730); carbamates – linear (907); carbamates – cyclic (354); carboxylic acids (4116); esters (8268); ethers (8736); hydroxypyridines, hydroxytriazoles, and all other hydroxyl heteraryls (653); imides – cyclic (28); ketones (6097); lactams (5510); lactones (935); nitrile (5299); phenols (1308); pyridines/pyrimidinones and all other cyclic aryl lactams (6637); sulfonamides (10198); sulfones (2916); sulfoxide (97); thioethers (2736); ureas – fully cyclic (524); ureas-linear (4007).

Despite the similarity cutoff of 0.99 between the Stage 1 and Stage 6 libraries (Figure 2), we determined that the effect of other similarity processing led to 86% and 96%, respectively, of the Stage 6 library being less than 85% and 90% similar to the Stage 1 library.

In the very final analysis, we undertook a series of random exports of 1000 compounds to scrutinize structures individually, locating occasional compounds that had hitherto escaped the functional group filters and modifying the filters accordingly. Ultimately, the quality of compound structures in an HTS library can really only be assessed upon independent scrutiny by medicinal chemists. For this reason, we have included in the SI the structures of one of these random exports of 1000 compounds (see Figure S2, pp 5–54).

The nature of our filters would of course remove most natural products. We see these and certain other types of rejected compounds as serving very useful roles but in segregated focus libraries and separate from the general screening libraries.

Table 5. Selected Individual Functional Group Filters in Filter 1 (see the SI, pp 55–59)

Entry	Filter	Category	Reasoning
1	 activated carbamate  electrophilic pyrimidine	A. Highly electrophilic.	High reactivity potential with protein nucleophiles and hence non-specific activity. Reactive enough to hydrolyse in HTS sample over time and not sensible to import & register into a screening database.
2	 oxonium  carbamic acid	B. Nonsensical	Likely to be so unstable that assay sample will not match structure and so it is not sensible to import & register into a screening database.
3	 peroxide  epoxide	C. Less electrophilic; may even be valid in a drug.	Reactivity potential with protein nucleophiles high enough to be problematic in a screening hit: likely to follow warhead-driven SAR; but may be suitable for inclusion in focus libraries.
4	  cyclohexadienes	D. Unstable.	Likely to become a more stable and predictable by-product in HTS sample, which won't be problematic <i>per se</i> , but not sensible to import such uncertainty or database inaccuracy.
5	  excessively halogenated groups CCl ₃ carboxylic acid CO ₂ H < 3 ester CO ₂ R < 3 nitrile CN < 3 thioether CSC < 3 iodine < 1 bromine Br < 2 chlorine Cl < 4 fluorine F < 6 furan < 2 thiophene < 2	E. Unnecessary starting point, even though may be valid in a drug.	Select simpler compounds first. If necessary, more elaborate analogues can be captured through "SAR by catalogue"
6	B boron, boronic acids etc NO ₂ nitro P phosphorous, phosphates etc Si silicon	F. Unattractive starting point even though may be valid in a drug.	These groups can be introduced later, but it may be hard to convince medicinal chemists to start with these compounds.

DISCUSSION

The most striking result of our file processing was a diminishingly small number of available compounds that met our selection criteria. In fact, as the ten vendors chosen represent about 80% of available lead-like compounds⁴ and as the Stage 1 library did not significantly influence compound selection, one could argue that the world's available lead-like HTS compounds can be represented by a library of just 341K compounds.¹⁶

Complementary observations can be inferred from other studies,^{4,17–21} and it is clear that there is a stark difference in the depth of available screening chemistries and that of theoretically possible screening chemistries^{22,23} (see the SI, p 92, for further discussion on this issue).

Ultimately, to what extent one agrees with estimations of the numbers of available lead-like compounds depends on to what extent one agrees with the stringency of the filters that are applied, in particular in our case, the functional group, PAINS, lead-likeness, and similarity filters. In the typical example shown in Figure 1, for example, these filters were responsible for removing 27.5%, 7.5%, 52.5%, and 10.7%, respectively, of the vendor collection. As shown in Figure 2, of the collected vendor files, similarity filters removed a further 73K out of 208K compounds (35%) and functional group filters a further 41K out of 135K compounds (30%). The notion of physicochemical lead-likeness (lower MW, fewer HBA and HBD etc.) is well

established and well discussed,^{4,17,19,25,26} and while our own criteria may differ in the detail, we are comfortable that this particular aspect need no further discussion. Conversely, further discussion is merited on the use of our functional group, PAINS, and similarity filters.

Functional Group Filters. Our functional group filters represent an imposing 569 functional groups (147 for Filter 1, 38 for Filter 2, 304 for Filter 3, 80 for Filter 4). Note that these are in addition to the 480 PAINS substructure filters that will be discussed separately, bringing the total filter number to more than 1000.

Those familiar with developing filters will understand that the output of every filter needs to be thoroughly checked to ensure that the intended compounds are being appropriately recognized and to the exclusion of unintended compounds. For example, a filter N–S to remove all single nitrogen bonds to sulfur will also remove all sulfonamides unless the filter is more precisely defined (using sybyl line notation) as N–S[TAC=2], where TAC stands for total atom count (on the sulfur atom). This may come as a surprise to the uninitiated.²⁷ Likewise, an imine filter simplistically defined as C=N could potentially remove extremely large numbers of drug-like heterocycles unless specified to not be in a ring: C=N[!r]. Even here, this will also remove all hydrazones unless the definition is revised. It is not always clear in lists of functional group filters whether functional groups are defined as intended.¹⁹ For our battery of

files, every functional group filter has been rigorously checked in this way.

Functional groups may be viewed unfavorably for several reasons. For example, because of their association with poor PK behavior or toxicity – the two often being associated with one another – and Pfizer,²⁸ Abbott,²⁹ Xiphora,³⁰ Boehringer,³¹ and more recently Merck³² have all published in this area. Other publications by AMGEN,⁶ GSK,⁵ Vertex,⁷ Abbott,^{33,34} Pyxis,¹⁷ BMS,³⁵ and ourselves^{9,36} have focused more on problematic functional groups in the context of screening, where reactivity is emphasized but where a variety of other possible reasons associated with poor optimizability is acknowledged.^{5,7,9,17,36,37} In only rare cases are more comprehensive listings of filters provided in machine-recognizable form,^{5,9,19,20,33,35} and even here it may not be obvious how context-dependent each filter is nor how stringently it should be applied for the alternative purposes of an independent research group. Confusion is exacerbated in some cases by unclear nomenclature. For example, hydrazides, which are acylated hydrazines by definition, have been classified as reactive⁵ and have become incorrectly known in some publications^{5,31,38} as acyl hydrazides. Also, hydrazones have been referred to as hydrazines,³⁹ but their respective reactivities may be very different. It is possible some confusion has resulted between hydrazides and/or hydrazones and the more reactive and hydrolytically labile acylhydrazones. Depending upon how the respective filters are precisely defined, this could result in capture or loss of unintended compounds.

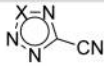
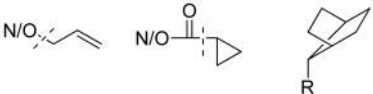
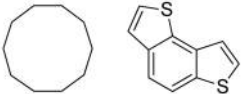
Workers at Johnson and Johnson have noted⁴⁰ that the question “should we buy this compound?” can have an answer that differs from “should we screen this compound that is in our inventory?” and differs again from “should we follow up on this screening hit?”. Flexible demerit systems have been implemented to help make this inherently subjective process as objective as possible.⁷ This is particularly useful for unattractive compounds that are already in collections but are not so bad that they need to be purged from the library. In our case, we had the luxury of importing a new library expansion from scratch, and so we did not seek to employ a demerit system but simply excluded any compound identified by the filters we developed. Our quest was clear: to obtain all possible compounds of the highest quality from a broad coverage of vendors before considering those of a lesser quality.

One unarguable reason to exclude compounds from purchase is through the evidence-based link between electrophilicity, protein reactivity, and promiscuous assay signaling.^{9,33,35} Interestingly, electrophiles that are most problematic in screening are not necessarily those that are most reactive in a conventional synthetic sense, while, conversely, some highly electrophilic synthons may not give rise to assay interference for reasons that are hard to understand.^{9,33,39,41} It may simply be that prior hydrolysis in screening stock to more stable compounds is involved, but regardless it is clearly inappropriate to have potentially electrophilic moieties in screening libraries. During our Stage 6 library expansion we observed numerous new electrophilic chemistries that escaped our PAINS filters and Stage 5 functional group filters, requiring new functional group filters to be defined. However, there have also been several other reasons why we have excluded certain compounds from our Stage 6 expansion. In the interests of transparency we therefore elaborate here our thought processes in developing our extensive functional groups filters and that have allowed for selection of the highest quality screening compounds.

In Table 5 we summarize six clear reasons to exclude certain compound types. First, and as just discussed, compounds that are sufficiently electrophilic to indiscriminately react with nucleophiles are clearly inappropriate. This includes obvious categories such as acid chlorides and well-known Michael acceptors such as α,β -unsaturated ketones. These are inevitably listed in all published functional group filters and so will not be commented further upon here. However, there are scores of other reactive groups that often remain less well-defined, in particular where the parent functional group, such as carbamate or sulfone,⁴¹ may be benign in some systems but become highly activated in others (entry 1). Our filters encode for the many dozens of the possible reactive variants of such otherwise benign functional groups (see the SI pp 55–87). Second, there are those compounds that perhaps have been accidentally or incorrectly incorporated in files selected for processing, such as oxonium and carbamic acid analogues (entry 2), or pentavalent carbon. These may not be specified in published filters and may only be noticed upon exhaustive scrolling through the many thousands of compounds retained at any given time in the filtering process. Third, there is a category of functional groups that are sufficiently less reactive that they may be viable in drugs, but whose reactivity nevertheless precludes their consideration for a general screening library of lead-like compounds: a target-based epoxide screening hit with a low micromolar IC₅₀, for example, is likely to follow warhead-driven SAR and be difficult if not impossible to optimize toward a selective lead with therapeutic potential (entry 3). Yet carfilzomib is a natural product-derived⁴² anticancer agent recently approved by the FDA that contains an electrophilic epoxide. Fourthly, there are compounds such as cyclohexadienes (entry 4) and numerous related compounds that are likely to oxidize on storage. The resulting aromatized counterpart would be stable and unlikely to be a nuisance screening compound *per se*, but it is simply not sensible to import compounds where the physical sample does not match the database definition. Fifthly, there are large numbers of compounds that are not necessarily reactive and are not assay interference compounds, but which we regard as simply unnecessary to include first-up in a screening library. These compounds may contain excessively halogenated rings or several unattractive groups such as three thioether linkages (entry 5). While such definitions potentially capture increasingly poor PK optimizability, our main driver was the principle of leadlikeness – that simpler analogues were generally always available and we would prefer to purchase those in preference. Finally, a sixth category comprised groups that might have unique properties but have been traditionally regarded as unattractive from a medicinal chemistry perspective and met with reticence by medicinal chemists. These include boronic acids, silicon-containing compounds, nitro groups. We note that progression of such compounds is certainly not without precedence (for example: trifluoromethylsulfonyl to favorably mimic nitro present in a screening hit;⁴³ there is emerging interest in boron-containing medicinal chemistry;⁴⁴ and silicon can usefully replace carbon as a carbon atom isostere⁴⁵), and we are actively reviewing some of these filters. However, it is also the case that reticence by medicinal chemists to engage with some type of compounds can represent a barrier to progression sufficient to exclude these compounds in the first place.

Exhaustive scrutiny of compounds remaining after application of Filter 1 allowed for identification of additional filters. Subsequent filtering through Filter 2 and then Filter 3 generally

Table 6. Selected Individual Functional Group Filters in Filter 3

Entry	Filter	Category
1	 X=any heavy atom	A/C
2	 N/O-CH ₂ -CH=CH ₂ N/O-C(=O)-Cyclopropyl R- numerous other fused carbocycles	E.
3		F.
4	carboxylic acid CO ₂ H <2 ester CO ₂ R <2 bromine Br < 1 chlorine Cl < 3 fluorine F < 4 thiocarbonyl < 1 sulfonamide/sulfone < 2 oxime < 2 ketone < 2 basic amine < 2 (unless piperazine) alcohols < 3 linear amides < 3 no more than 1 thioether, hydrazone, hydrazide, oxime, ketone, ester, acetal, hemiaminal, amidine/guanidine, phenol, in the same molecule N+O < 7	G.

represented progressive harshening of criteria. However, these are hardly controversial, and underlying reasoning can be clearly identifiable to be the same as those outlined in Table 5. For example, shown in Table 6 (entry 1) is one of many nitrile-containing moieties that we defined because our collective experience suggested that these would be problematically reactive, a decision supported by subsequently located literature.⁴⁶

It was also the case that certain simple groups were becoming increasingly represented the more we filtered out others, for example, allyl and cyclopropyl amines/amides/ethers and numerous assorted fused hydrocarbons (Table 6, entry 2), large macrocycles, and multiply fused aromatics (Table 6, entry 3). We were comfortable in removing large macrocycles and multiply fused aromatics as these represent unattractive starting points for drug discovery. We were also comfortable in removing allyl and cyclopropyl amines/amides/ethers because we determined that simpler counterparts were invariably present (e.g., *N*-Et, -NC(=O)Et, *R*-O-Et).

At this stage, we also applied more stringent limitations on the degree of functional group and heteroatom multiplicity allowed per molecule (entry 4, Table 6). Once again, we do not see these criteria as being particularly controversial because our aim to retain only the most lead-like compounds.

Indeed, we deliberately have very seldom excluded compounds on the sole basis that they contain a functional group associated with problematic PK properties or toxicology. For example, we know of some drug discovery companies that exclude hydrazones from screening libraries on the basis of a link to poor PK. As long as that group is not associated with assay interference, the potential for downstream liabilities in a given screening hit does not particularly concern us, especially if the group in question might represent linkage of binding elements in a manner not otherwise represented. In the case of

a hydrazone, this is quite possible, and in fact such a compound was the basis for one of our more successful medicinal chemistry campaigns involving the Bcl-x_L-Bim protein–protein interaction (manuscript in preparation). In such cases our approach is to engineer-out a potentially problematic group during optimization, if it is determined that the group in question is indeed likely to be a liability. Thus in our secondary selection set (Figure 2) we deliberately allowed for inclusion of groups such as acetals, aminothiazoles, chromones, hydantoin, hydrazones (not acylhydrazones), oximes, and thioethers. In some settings^{17,19} these might be excluded from purchase or at least receive demerit points, but we have no evidence that these are problematic in a screening sense and so have not excluded them: any that are so functionalized to become PAINS will have been removed in the PAINS filters. However, as shown in Table 6 (entry 4), we have still limited the number of such potentially liable groups to just 1 per molecule in order to remove those compounds likely to be least progressable.

The random export of 1000 compounds (SI, pp 5–54, Figure S2) comprises an excellent data set to exemplify the nuances of functional group exclusion or inclusion and to what degree our filters could be considered relatively lenient. For example, there are around three dozen compounds in the random set of 1000 that could be considered the least attractive. As there is a paucity in the literature on the reasoning behind functional group assessment, it is apposite to the academic drug discovery community for our thought processes to be elaborated on, and so we select 13 for discussion as shown in Figure 3.

For example, it was anticipated that 1, 4, 5, 6, 12, and 13 might be problematically electrophilic. Reasonable numbers of analogues of 1, 4, 5, and 12 containing the substructure shown were present in the Stage 1 library (69, 605, 6, and 25,

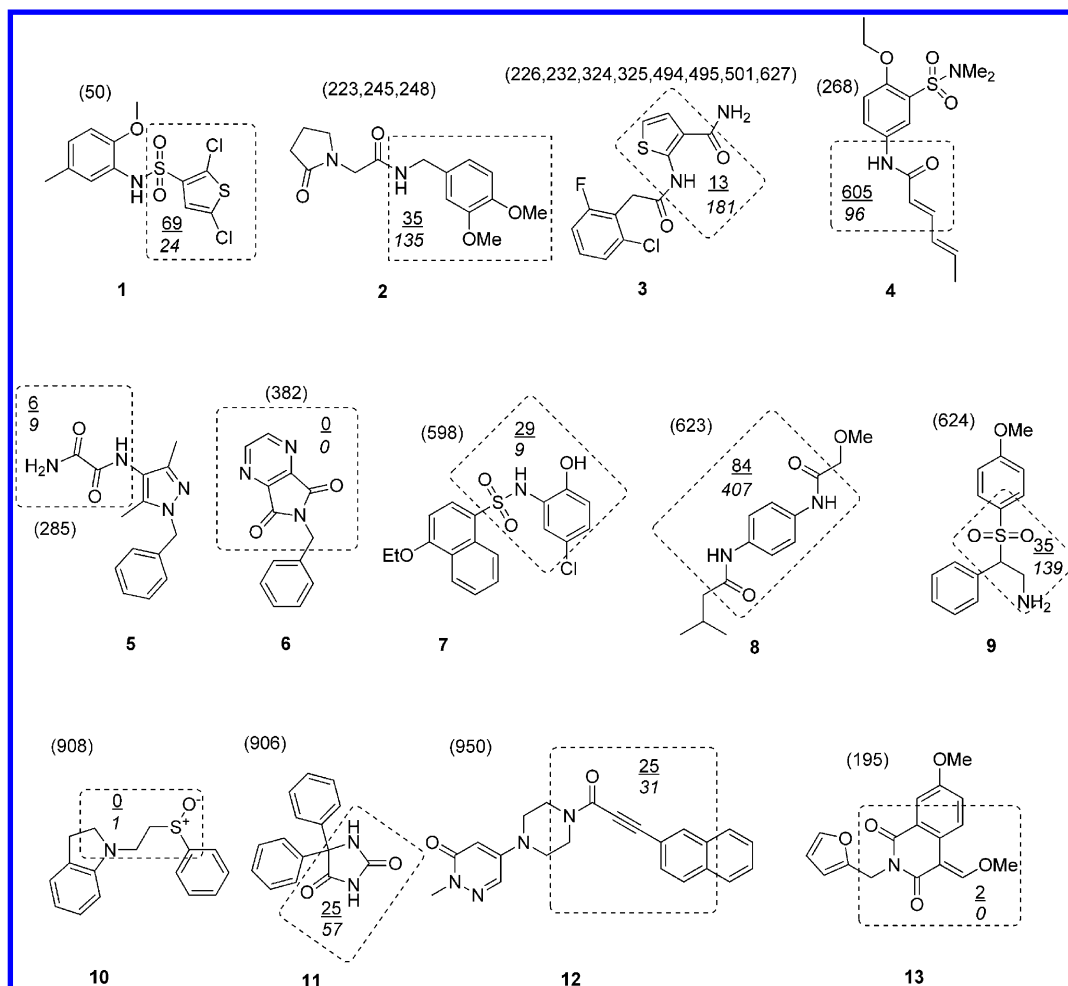


Figure 3. Some of the few compounds that are less attractive in a random export of 1000 in the penultimate version of the Stage 6 library. The numbers in parentheses refers to the relevant entries in Figure S2. The numbers underscored and italicized refer to the number of compounds in the Stage 1 and Stage 6 libraries respectively that contain the substructure within the dotted box.

respectively) such that we could undertake analysis for PAINS behavior using our previously defined criteria.⁹ No such behavior was evident. Compounds 6 and 13 were absent or scarce in the Stage 1 library but were deemed to be sufficiently electrophilic that we introduced new filters to remove these from the Stage 6 library just prior to final purchase. Compound 2 represents an electron-rich aryl that as just discussed is not excluded purely on the basis of possibly poor PK: first, such compounds could still be useful tools; second, we would seek to engineer out poor PK if an issue; third, there are diseases where intravenous administration is relevant and where poor PK due to first pass clearance is less so; fourth, there are many drugs in the market and in development with electron-rich aryl rings. Compound 3 represents one of several 2-amino-3-carbonylthiophene derivatives. We do not particularly like these compounds as they are known to be redox active³³ and some are also PAINS. Nevertheless, these specific ones passed our PAINS substructure filters, and we also are conscious that there are examples where such screening hits have progressed. For example, AstraZeneca started with such a compound⁴⁷ and subjected it successfully to a regioisomeric switch to a less problematic 3-amino-2-carbonyl system⁴⁸ that appears to be progressing through phase II clinical trials. We have therefore chosen to retain these particular compounds but will watch their performance with interest and revise our future PAINS

filters accordingly if necessary, especially as these compounds number some 181 members in the Stage 6 library.

Compounds 7 and 8 are unattractive in the sense that they are hydroquinone-like and there would be concern that facile oxidation would lead to reactive quinoid-like compounds: many similar compounds are PAINS.⁹ Nevertheless, significant numbers of compounds in the Stage 1 library (29 and 84, respectively) with the specific substructures shown were not problematic and so such compounds have been allowed in our Stage 6 library.

The β -aminosulfone 9 and the β -aminosulfoxide 10 might be considered to potentially undergo retro-Michael degradation to reactive vinyl species analogously to β -aminoketones as we have discussed previously.³⁶ However, even though the sulfone group is intrinsically more electron-withdrawing than a ketone group, electronic and geometrical factors dictate a lower reactivity for the former in their respective vinyl counterparts.⁴⁹ While reactivity can vary greatly depending on the nature of substituents, it is not clear that degradation of β -aminosulfones to vinylsulfones is a problem in HTS libraries nor that the latter, even though it has been used as a soft warhead, would be nonspecifically problematic.^{39,50} This is supported by our analysis of the screening history of the 35 β -aminosulfones (all *N,N*-dialkyl) that were present in our inaugural Stage 1 library that revealed these to be extremely clean (enrichment factor⁹

≪ 10%). We have therefore allowed such compounds in our Stage 6 HTS library.

We show hydantoin **11** in this section because we have often been in the presence of industry-experienced medicinal chemists where these are dismissed and so are correspondingly often excluded from screening libraries.¹⁹ It is hard to find the justification of this other than some unfortunate toxicological history with drugs such as phenytoin (**11** is actually phenytoin).²⁸ Hydantoins can provide rich functionality in a small volume and as screening and tool compounds, our current status is to view them favorably, at least as starting points for optimization.

We were intrigued that numbers of analogues for some substructures differed greatly between Stage 1 and Stage 6 libraries. For example, analogues with substructures shown for **2** (135), **3** (181), **8** (407), and **9** (139) are vastly more numerous in the Stage 6 library. We determined that vendor A was the major contributor of these (61%, 68%, 58%, and 69%, respectively). Given that vendor A supplied 58% of the compounds for our Stage 6 library, this may not seem so surprising. However, in some ways it is because it implies that vendor A's contribution is not necessarily through new diversity space but, at least in part, by greater representation of certain compound classes. Conversely, the substructures defined in **1**, **4**, and **7** are relatively underrepresented in the Stage 6 library, presumably because these are either less represented in vendor collections or remaining analogues are too similar to each other and so are removed by the similarity filtering.

In summary, in the interests of learning more about some less attractive compounds that are not discussed in the screening literature, we have allowed a small number of carefully considered compounds to be included in our secondary selection set. In most cases such compounds may be potential electrophiles. We will track the behavior of these few probes with interest, especially involving targets with reactive site side chain nucleophiles. In this context, it would be highly useful to the academic drug discovery community for pharma to release what is almost certainly a great depth of data on functional group behavior accumulated over years of HTS. This could improve the objectivity of functional group exclusion and screening hit scoring.

It is beyond the scope of this manuscript to detail here our consideration for each and every functional group filter, but in most of the cases that warrant comments such information is provided within the parent filter file (see the SI).

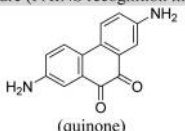
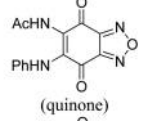
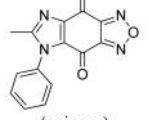
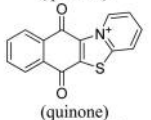
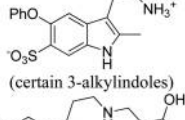
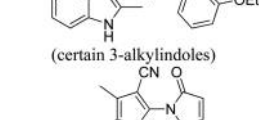
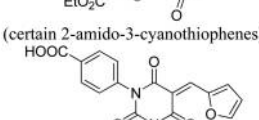
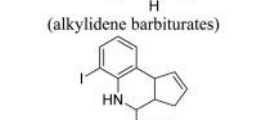
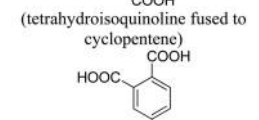
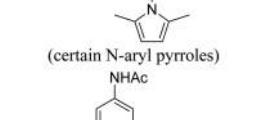
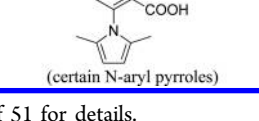
PAINS Filters. Another process that merits discussion is the use of filters to remove all PAINS,⁹ which from Table 2 can be seen to amount to some 162K compounds. These are nuisance compounds that in our experience have inevitably led to dead ends. We are therefore comfortable in excluding these, especially as by their inclusion, our processing protocol may lead to their selection ahead of far more attractive compounds that may be available. These compounds are particularly subversive as their early biological data may be compelling and it may take effort to convince others not to progress them, and we have found that they routinely pass as "confirmed hits". As we have noted before, the appearance of these compounds in hit sets should be accompanied by the recognition that they are highly likely to be hits for all the wrong reasons and the chance of them representing "real" and optimizable compounds is very small relative to the chance that they represent a promiscuous and poorly optimizable starting point. By way of illustration, we recently published a high throughput assay for inhibitors of

MOZ, a histone acetyl transferase (HAT), where only one real hit was unearthed.⁵¹ We did not disclose the structure of this compound since genuine inhibitors of HATs are rare and highly sought.^{52–54} One of our HTS libraries prosecuted by this screen was our Stage 1 library (that still contains PAINS), and, as we routinely do with hit sets from this library, we applied the PAINS filters to the primary hit set. This removed 180 of 527 compounds that had been selected up to that point of triage. Ordinarily we would not further investigate the removed PAINS, but for the purposes of the research discussed herein we have titrated the 52 PAINS that were relatively clean: in other words, those that had registered as hits in only 3 of the 6 screens that had originally been selected for the development of our PAINS filters.⁹ Of these 52 PAINS, 21 tested as inactive and 20 as nonspecific in that they were active in the counterscreen. This left 11 PAINS that registered as selective hits for MOZ. These are shown in Table 7.

These 11 PAINS represent 6 typical classes that we frequently see as HTS hits, these being quinones (**14–17**), certain alkyl indoles (**18**, **19**), certain 2-amido-3-cyanothiophenes (**20**), alkylidene barbiturates (**21**), cycloalkenyltetrahydroisoquinolines (**22**), and certain N-aryl pyrroles (**23**, **24**). Of note is the concomitant presence of potentially contributing interference moieties, such electron-donating groups in quinones (**14** and **15**), furazans in quinones (**15** and **16**), pyridinium in quinones (**17**), and maleimide in certain thiophenes (**20**). That a maleimide is present in the first place is a reflection of the very modest functional group filtering deployed in establishing our Stage 1 library. These subgroups may all contribute to unusual spectroscopic and redox or electrophilic reactivity in ways that in some cases remain to be precisely determined as we have commented elsewhere.^{9,36,37} We note that since our previous discussion³⁶ on compounds such as **22**, it has come to light that yet another interference mechanism involving heavy metal contamination may be at play (Gadolinium in this case), with the carboxylic acid playing a participatory role in chelation (Uli Schmitz, personal communication). The presence of more than 1 member for a number of the classes represented – there are two 3-alkyl indoles and two aryl pyrroles for instance – may be readily misinterpreted as SAR rather than a hallmark of a problematic core. Moreover, PAINS can furnish striking data. For example, aryl pyrroles such as **23** returned an IC₅₀ inhibitory value against MOZ of 0.51 μM in this assay (which was designed to pick up competitive inhibitors of the substrate, AcCoA), and there was no measurable activity in the counterscreen. Bis carboxylic acids are known phosphate isosteres,⁵⁵ and one might reason that this makes perfect sense if it is to mimic the substrate Ac-CoA. Combined with a structure that to the uninitiated may appear acceptably lead-like, this compound might be earmarked for medicinal chemistry optimization and ranked higher than the real MOZ inhibitor discovered in this assay.⁵¹ Our experience suggests that this would be a mistake, as these very same compounds have previously misled us in other screening projects. We have discussed these types of compounds elsewhere.^{9,36}

Although absent from Table 7, by far the most obviously problematic PAINS class are the alkylidene rhodanines, but even here there is continuing debate in the academic community whether these are good¹¹ or bad.¹² We are even more concerned about more subtly subversive PAINS such as the alkylidene pyrazolones that as we have discussed³⁶ are routinely being progressed as valuable tool compounds and

Table 7. PAINS That Were Active against MOZ after an HTS Campaign and That Appeared to Be Selective

Cpd	Structure (PAINS recognition moiety)	IC ₅₀ (MOZ), μ M	IC ₅₀ (Counter) ^a
14	 (quinone)	1.3	>62
15	 (quinone)	4.5	>62
16	 (quinone)	75	>125
17	 (quinone)	11.1	>62
18	 (certain 3-alkylindoles)	4.8	>62
19	 (certain 3-alkylindoles)	22	>125
20	 (certain 2-amido-3-cyanothiophenes)	11.7	>62
21	 (alkylidene barbiturates)	20	>125
22	 (tetrahydroisoquinoline fused to cyclopentene)	24	>125
23	 (certain N-aryl pyrroles)	0.51	>62
24	 (certain N-aryl pyrroles)	7.2	>62

^aSee ref 51 for details.

being highly cited with no consideration of potential PAINS behavior. A most compelling reason to be concerned about these is highlighted in a recent publication,⁵⁶ where complex protein labeling behavior of alkylidene pyrazolone SJ-172550 has rendered it delisted as a progressable lead compound. Regrettably, high impact journals continue to portray PAINS as useful tool compounds,⁵⁷ but this also illustrates how on some biological data alone they can be difficult to detect. As we have previously discussed,⁹ some PAINS can form noncovalent

complexes with proteins, and so one could choose to include them in screening libraries and install a demerit system instead. However, to purchase such compounds ahead of others makes no sense to us. Also as we have previously discussed,³⁶ one wants the best possible starting point in drug discovery and the chances that a PAINS hit represents a progressable compound are vanishingly small compared with the chances that it does not. Furthermore, such compounds may provide cell-based readouts that appear to be linked to the targeted mechanism-of-action but are not.^{9,58} It may be extremely hard to convince colleagues to not pursue such compounds, and we have adopted the stance that it is better not to purchase such compounds in the first place. For this reason we have made all possible attempts to make PAINS filters readily accessible.^{9,36,59,60} The only conceivable drawback to excluding PAINS comes from a purely academic perspective because there is much to be learned about PAINS and their interference mechanisms. Cumulative screening campaigns and follow-up mechanisms of action studies could provide useful information in this regard, but this knowledge cannot be gleaned if those compounds are not present in the first place.

Similarity Filters. Finally, we now discuss the removal of excessively similar compounds as measured by the Tanimoto coefficient, the effects of which as shown in Figures 1 and 2 are profound in terms of the percentage of remaining compounds removed, in agreement with the observations of others.¹⁸ In a seminal paper by Verheij¹⁷ in 2006, analysis of commercial libraries (using a scaffold classification, nonfingerprint approach) revealed vendor-dependent average cluster sizes of 2–11 similar analogues, but cluster sizes of up to 450 analogues could be identified. The percentage of singletons per library varied from 47% to 74% when calculated relative to the total number of clusters per library. For example, the ChemDiv catalogue contained 9,921 singletons (61.6%) out of a total of 67,446 compounds that contained 6,172 (nonsingleton) clusters. In the 2010 study by Chuprina et al.⁴ it was observed that 93.7% of the 2 M unique and drug-like commercially available compounds had a similarity of between 0.80 to 0.99. Analysis of these commercial libraries (using a sphere exclusion, fingerprint approach) revealed vendor-dependent average cluster sizes analogues more than 80% similar to each other varying from 3.7 for one vendor to 26.7 to another. For the latter vendor with a library size of 196,064 compounds, there were some 372 clusters containing more than 100 analogues! The percentage of singletons for all vendors averaged 41% when calculated relative to the total number of clusters per library (324,905 out of 5,183,506 total compounds or 6.3%). In general, the analytical results of these two studies 4 years apart using slightly different methods are remarkably concordant. The 2008 study¹⁹ by Brenk et al. of 222,552 lead-like compounds from around ten vendors led to the identification of 31,105 singletons (14% of total compounds) and 89,245 compounds in 9,705 clusters comprising compounds within the similarity range 0.80–0.90. The more stringent clustering criteria led to singletons comprising 76% of the total number of clusters (31105 + 9705). Despite the cutoff of 0.90, although clusters were greatly reduced in size, some still comprised up to 60 analogues.

Assuming at least an approximate correlation between the concepts of (nonsingleton) clusters and compound class, a significant interpretation from the above studies is that there are remarkably few different types of classes in available chemical space, perhaps in the order of only 20,000 or so with

the rest comprising analogues. Interestingly, recent scaffold-based analyses lead to similar conclusions.⁶¹

The very large attrition rate that we have observed using a cutoff of 0.90 in Table 3 is now understandable and one can conclude from these studies that the net result of our similarity filtering should be to allow the influx of relatively large numbers of singletons and small analogue sets while disallowing the presence of clusters that are fewer in number but excessively populated with similar compounds. This is precisely what we observe, and analysis of the Stage 6 library reveals that it comprises 76,177 singletons, or 67% of the total number of compounds, using the typical definition that any compound less than 80% similar to any other (i.e., cutoff 0.80) is a singleton.^{4,19} Of these, 24,121 compounds become partnered in the combined library set, and so 52,156 singletons are effectively contributed by the Stage 6 library. The remaining 35,823 compounds in the Stage 6 library are between 80% and 85% similar to each other and belong to clusters of between 2 and ~15 compounds (note that the groups of 24,121 and 35,823 compounds can be up to 90% similar to existing compounds and up to 99% similar to the almost-exhausted inaugural Stage 1 HTS library).

The question naturally arises, in removing such large numbers of compounds deemed to be too similar to each other for inclusion, are we nevertheless losing useful compounds – those that may provide a hit where no other similar included analogue does.

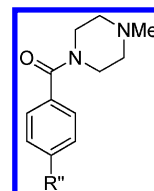
A germane 2002 study by Martin et al. concluded that the common policy of pharmaceutical companies to exclude from purchase available compounds more than 85% similar to those already in the screening deck, on the basis that any relevant biological activity would be captured in the existing analogues, was flawed.¹⁸ Rather, a compound more than 85% similar to any given active had only a 30% chance itself of being picked up as an active in the same assay.¹⁸ In other words, as noted by Martin, if only one compound of a similarity set is screened, there is a 70% chance that the activity within this cluster will not be discovered. Martin went on to note that such sharp SAR is not at all at odds with the experience of medicinal chemists yet sits paradoxically with the similarity-based exclusion-for-purchase philosophy. On this basis, we deemed it possible that our similarity-exclusion processes might filter out potentially unique screening hits. However, there remained some uncertainty about this because of inherent differences in the nature of the Martin study compared with ours. From a medicinal chemistry perspective, we were interested in tracking a putative SAR set during similarity filtering in order to better understand its effects in a more meaningful fashion for chemists. For this purpose, we selected *p*-substituted benzamides of *N*'-methylpiperazine as a group of synthetically accessible and drug-like compounds likely to be well represented in vendor catalogues, focusing on vendor B.

First, it is instructive to summarize the general attrition observed for vendor B during processing. In Table 3 (entry 1), it is seen that vendor B has 161,726 compounds left to select from after the extensive processing of the vendor catalogue as shown in Table 1. We now select all compounds from the remaining vendor B compounds that are no more than 90% similar to those already present in the 301,412-strong reference database (entry 2), which comprises the extant reference database of 180,759 summed with the 120,653 compounds selected just prior from vendor A through analogous processes (entry 4). This results in selection of only 49,090 of vendor B

compounds from the 161,726 available. In other words, the Tanimoto coefficient judges there to already be large numbers of similar compounds already existing or selected for purchase, as just discussed.

We now look at the passage through these processes of all *N*'-methyl-*N*-piperazinylbenzamides that are available from vendor B immediately prior to similarity exclusion. To simplify the process we look just at those compounds with a simple para substituent. As shown in Table 8, such compounds comprise a

Table 8. *N*-Me Benzoyl Piperazines with a Simple Para Substituent That Are Available from Vendor B

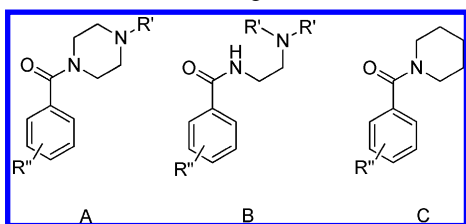


cpd	R''	present after 90% similarity removal?	entry in Table 9 responsible for removal
25	F	N	49, 50, 62, 69
26	Cl	N	47, 67, 70, 74
27	NHCO- <i>n</i> -Bu	N	71
28	NHSO ₂ Me	N	55
29	NHSO ₂ Et	N	55
30	NMe ₂	N	60, 65
31	piperidin-1-yl	N	61
32	CH ₂ -pyrrolidin-1-yl	Y	-
33	O- <i>n</i> -Pr	N	52, 56
34	O- <i>n</i> -Bu	N	56
35	O- <i>i</i> -Bu	N	56
36	OBn	N	52, 73
37	Me	N	45, 46, 48, 53, 63, 72
38	Et	N	57, 72
39	Pr	N	58, 72
40	Bu	N	58, 72
41	<i>t</i> -Bu	N	68, 72
42	Ph	N	46, 51, 54
43	SO ₂ NPr ₂	N	59, 64, 66
44	SO ₂ -piperidin-1-yl	N	59, 64

relatively good SAR set, numbering some 20 analogues. After performing the operation just discussed where only those compounds are selected that are less than 90% similar to those already present in our collection, quite remarkably – and underscoring the severity of the effects of similarity exclusion and confirming our expectations that using vendor B would result in significant attrition – only 1 compound (32) is selected for purchase.

We then investigated what compounds in our current collection were assessed to be more than 90% similar to those in Table 8 and that were therefore responsible for the lack of selection of the latter. These 30 compounds, which arose from one of Stage 3, Stage 4, or Stage 5 HTS libraries, are shown in Table 9.

Thus, compound 25 (Table 8) was not selected because four compounds were already present that were judged to be more than 90% similar to it, these being 49, 50, 62, and 69 in Table 9. It may be unexpected that compounds 50 and 62 are regarded as being so similar to 25 because they contain large

Table 9. Existing Compounds Judged To Be More than 90% Similar^a to the Available Analogues for Selection in Table 8

cpd	class	R'	R''
45	A	H	H
46	A	Bn	4-Me
47	A	Bn	4-Cl
48	A	COPh-3-Me	3-Me
49	A	H	4-F
50	A	3-FBn	4-F
51	B	Et	4-Ph
52	A	4-EtOBn	4-OMe
53	A	3-MeBn	H
54	B	Me	4-Ph
55	A	Bn	4-NHSO ₂ Et
56	A	Me	4-O-iBu
57	A	Allyl	4-Et
58	A	n-Pr	4-n-Bu
59	B	Me	4-SO ₂ -(piperidin-1-yl)
60	A	Bn	4-NMe ₂
61	C	-	4-NEt ₂
62	A	COPh-4-F	4-F
63	A	COPh	H
64	A	Bn	4-SO ₂ -(4-Me-piperidin-1-yl)
65	A	Me	4-NH ₂
66	A	Me	4-SO ₂ NPr ₂
67	A	Bn	3-Cl
68	B	Me	4-t-Bu
69	A	Me	3-F,4-Me
70	B	-CH ₂ CH ₂ N(Et)CH ₂ CH ₂ -	4-Cl
71	B	Me, Ph	4-(2-oxopiperidin-1-yl)
72	A	H	4-Et
73	A	Ac	3-OBn(3-Me)
74	A	i-Pr	4-Cl

^aSome compounds in this table are more than 90% similar to each other due to different similarity considerations established for Stage 2 and 3 libraries.

and distinct R' groups. As has been noted,¹⁸ this is an example where the fingerprint-derived Tanimoto coefficient does not necessarily match biologically relevant chemical similarity, and

it is easy to imagine 50 and 62 could be inactive against a given target because of steric reasons where 25 would be active. In contrast, compounds 49 and 69 would be regarded by most medicinal chemists as close analogues likely to have similar biological activity to 25 and so lack of selection of this compound would not necessarily pose a problem (to a medicinal chemist). All these compounds bear either a 4-F or 3-F suggesting its presence is a key factor that contributes to fingerprint similarity, whereas the N-piperazinyl substituent is largely irrelevant for these purposes.

On the other hand, compound 27 in Table 8 is not selected because it is assessed as being more than 90% similar to 71 in Table 9, yet there are sufficient differences between the two molecules such that a medicinal chemist could argue that 71 may be inactive for a given biological target but that 27 might not be inactive against this target. Compound 71 has a bulky phenyl group as an N-piperazinyl substituent and instead of a linear secondary amide as the R'' group in the 4-position of the phenyl ring, it has a lactam. Once again, the N-piperazinyl substituent does not seem to significantly influence the fingerprint definition, but an amide bond in the R'' group is clearly key. In other words, 27 is not adequately 'encoded for' in the screening deck, and we may consider to have effectively lost an active – and a future new class – due to an 'incorrect' classification that 27 is too similar to 71.

However, the situation is not as simple as this because just as the Tanimoto coefficient may regard medicinal chemistry dissimilar molecules as similar, so it may assess medicinal chemistry similars to be dissimilar. A substructure search illustrates this point. As shown in Figure 4, there are 234 compounds already in our HTS libraries that possess the substructure 75 that is also present in 27. Of these are included a number of compounds that from a medicinal chemistry perspective are very closely related to 27 such as 76–79 even though they are regarded by the Tanimoto coefficient as less than 90% similar (and so do not appear in Table 9). The pairwise similarity comparison is shown in Table 10.

Table 10. Pairwise Similarity Comparison of Analogue 27 with 76–79 as Assessed by the Tanimoto Coefficient Using Unity Fingerprints

	27	76	77	78	79
27	1.00				
76	0.86	1.00			
77	0.89	0.86	1.00		
78	0.76	0.79	0.71	1.00	
79	0.66	0.68	0.62	0.87	1.00

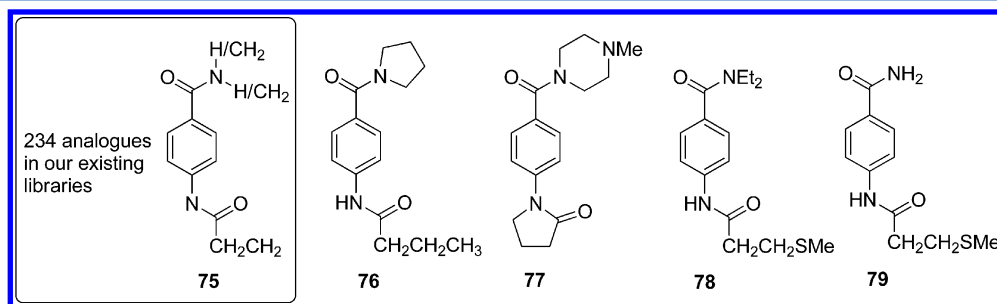
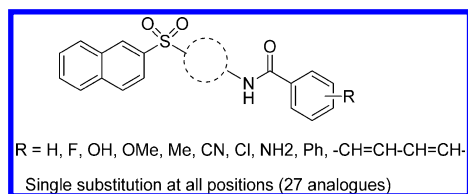
**Figure 4.** Compounds 76–79 are closely related to 27 and represent 4 of the 234 compounds that we currently have that possess the substructure 75.

Table 11. Effect of Similarity Removal on a Real SAR Set



scenario	number of compounds: commercially available/unavailable											
	no similarity removal				similarity removal (>0.85) ^a				similarity removal (>0.95) ^b			
	IC ₅₀ (μM)				IC ₅₀ (μM)				IC ₅₀ (μM)			
	<10	<30	<50	>125	<10	<30	<50	>125	<10	<30	<50	>125
A	1/2	4/2	0/2	0/16	1/1	0/1	0/0	0/0	1/1	1/2	0/0	0/0
B	0/2	4/2	0/2	0/16	0/0	0/2 ^c	0/0	0/0	0/1	1/2	0/0	0/0

^aAs described herein, involving several exclusions at 0.90 and a final exclusion with a Tanimoto coefficient cutoff of 0.85. ^bHypothetical situation where all similarity exclusions of 0.90 and 0.85 used in Figures 1 and 2 are replaced with an exclusion radius of 0.95. ^cBy one of the quirks of similarity filtering, removal of the original hit led to selection in Scenario B of a new analogue (3-F) in the <30 μM category which became responsible for exclusion of the 3-Ph (*t* = 0.91!) commercially unavailable active that had been present in Scenario A with an IC₅₀ of less than 10 μM.

A medicinal chemist could therefore argue that **27** is reasonably represented by **76–79** even though it was not excluded on this basis. Almost entirely analogous comments that have been made for selected examples **25** and **27** can be made for all other entries in Table 8 and their similarity analogues in Table 9. Therefore, it seems that despite the discrepancy between fingerprint-derived similars and medicinal chemistry similars, libraries established via excluding the former may still end up providing adequate coverage for highly populated classes. Support for this was gained through a retrospective substructure search for medicinal chemistry similars not captured through their fingerprints. We are not aware of this sort of illustrative example having been undertaken before and think it is particularly useful from a medicinal chemist's perspective.

But what if a particular class is not highly populated in available chemistry space? Our experience is that this is the case for most compound classes, as we have discussed herein. A pertinent example comes from a highly successful hit-to-lead program currently active in our laboratories, which began with a single hit from an HTS campaign designed to look for compounds with greater than 50% inhibitory activity at 10 μM. This compound required a naphth-2-yl-sulfonyl group linked to a substituted benzamide group as shown in Table 11. We discovered this class because a particular substituent on the phenyl ring confers activity such that it is identified using the 10 μM cutoff. In the course of elaborating SAR, we know the IC₅₀ values of all analogues designed to systematically probe the phenyl ring SAR. It is useful to consider the set of 27 simple analogues shown in Table 11 and the effect that similarity removal has on retention of actives and which of these are commercially available. Shown first is the hypothetical case where compounds are imported into our libraries without application of similarity filters and so all 27 analogues would be tested. Here (Scenario A), it can be seen that of the 27 analogues, 3 have an IC₅₀ of less than 10 μM and one of these (the original hit) was commercially available. Remaining analogues were less active, and indeed sixteen were inactive at the maximum concentration tested. Of note, 22 of the 27 analogues of this simple probe set are not available from vendors, and it is our experience that such relative paucity of SAR probing in available chemistry space is very common when one part of the molecule is relatively unusual (such as the

naphth-2-yl-sulfonyl group). We now look at our combined Stage 1–6 libraries after the similarity exclusion as already described has taken place.

Here, it can be seen that large numbers of analogues have been removed, and we have only one compound in our library with an IC₅₀ of less than 10 μM (the original hit) but that is sufficient to allow for the discovery this class of compounds. If all 27 analogues had been commercially available, we would have identified another screening hit in this category. There is a third analogue that would have been missed through employing a 10 μM cutoff and that is not commercially available. Were we to have been less stringent and only excluded compounds at the level of >95% similarity, two more actives would be present, one of these being commercially available and one not, and neither would be identified with a 10 μM cutoff.

This situation is precarious. Our whole medicinal chemistry campaign is predicated on a particular single analogue having been present in our library that met the 10 μM cutoff criterion. What if that particular analogue had not happened to have been made by any of the vendors? This scenario (B) is also shown in Table 11, where we have repeated the similarity exclusion simulation with our active compound removed. Here, it can be seen that with a final exclusion radius of 0.85, we would not have identified any compounds at the 10 μM cutoff, and this class would have remained undiscovered and our target protein without any known inhibitors.

There are three clear ways to increase the odds of discovery. The first is to substantially relax the similarity exclusion radius. In the case at hand with a 10 μM cutoff imposed, this represents a solution for Scenario B only if the single hit had been commercially available (which it is not) and there is no improvement in the odds for Scenario A. We have no doubt that for other cases the results may be more convincing, but it is also the case that such relaxation of the similarity exclusion radius we know would come at the cost of large increases in the library size (due largely to influx of analogues of existing classes rather than new classes), to the extent where neither the libraries nor the primary screening campaigns may be affordable in an academic setting.

The second clear way to increase the odds of hit discovery is for HTS libraries to contain more analogues for relatively under-represented compound classes. This is possible even with a similarity exclusion radius of 0.85 if sufficient analogues are

available from which to select. This is therefore dependent on vendors appropriately and systematically populating analogue space to a greater extent than is often currently the case. We imagine this would lead to an increase in library size of several-fold but to what extent is uncertain. In Scenario A and using our current similarity exclusion criteria, this would have doubled the number of hits identified, albeit just from one to two. In Scenario B, this would still be insufficient unless we introduce the third way of improving odds of hit discovery, which is to allow the identification of weaker compounds in the primary screen. In combination with better SAR representation, we believe it is this last approach that is the most compelling. It is our experience that in order to obtain an acceptably low hit rate, there is a natural tendency for screening personnel to lower the testing concentration as a means to achieve this aim rather than improving the assay signal-to-noise to allow a higher concentration of test compound, which may be more involved. We think this habit is fraught with danger and likely to lead to unsuccessful HTS campaigns for difficult targets. Rather, even though more work may be involved in the hit confirmation and triage process, we would recommend screening at higher concentrations in such cases. The challenge here is therefore improved technologies to triage larger and noisier primary hit sets to most efficiently remove false hits. Of course, the presence of PAINS could become problematic in such hit sets obtained at higher concentrations, unless they are absent in your library as they are for our Stage 4–6 HTS libraries. Medicinal chemists should be involved at all stages of an HTS campaign and especially in scrutinizing pilot screen hits because compound recognition can reveal assay flaws that can then be investigated. It is our perspective that there can be a tendency for screening groups to work independently from medicinal chemists and screen the full compound deck with a view to presenting the selected hits to the medicinal chemists for optimization.

Our strong views on how to maximize the chances of hit discovery for difficult targets are not drawn from this one case of anecdotal evidence. Rather, the example discussed nicely emulates our experience in most HTS campaigns, where for difficult targets one or two identified hits are shown to later be accompanied by more numerous weaker hits belonging to the same class that were already initially present in the screening deck. We have run screens on difficult targets where no hits were discovered and now are left wondering whether the outcome would have been different with a combination of higher screening concentrations and better elaboration of vendor analogue space that would have led to greater analogue representation in our HTS libraries.

In summary for this section on similarity filtering, some fundamentally important ramifications arise. First, medicinal chemists would disagree with a fingerprint-derived assessment that our Stage 6 library comprises 67% of singletons. We have undertaken a series of diverse substructure searches of so-called singletons invariably returned somewhere between 1 to 4 other SAR-relevant analogues within the Stage 6 library. For this reason, we think the view held among some that fingerprint-derived similarity singletons be excluded from HTS libraries on the basis that SAR will be absent, is a mistake because this may not necessarily be the case. Furthermore, singletons that are genuine medicinal chemistry singletons may occupy precious diversity space and in fact have underpinned some of our most successful medicinal chemistry campaigns. Additionally, student projects based on elaboration of interesting singletons may be

an entirely worthwhile pursuit in an academic setting. The second fundamental realization from our analysis was that it became clear to us that to screen only the Stage 6 library in isolation of the other libraries – in particular Stages 4 and 5 – is not sensible as there will be grossly inadequate coverage of diversity space.

CONCLUSIONS

In establishing our most recent Stage 6 library selected from vendors that have 80% coverage of available chemistry space, we disclose here our astonishing revelation at how relatively few compounds passed our functional group and similarity filters. Indeed, from this exercise we estimate that the globe's repository may be represented by a total of fewer than 350,000 of such compounds. We believe this is the first time such an evidence-based case with full data disclosure has been made for such a relative paucity of available compounds when filtered by robust lead-like criteria. Principal drivers for compound exclusion were physicochemical lead-likeness, and functional group and similarity filters, though PAINS filters also removed significant numbers of compounds. The concept of lead-likeness is well established and relatively objective, but reasons for functional group exclusion can be varied and more subtle. We have outlined six clear notions that we deployed as to why functional groups may be excluded from high quality screening libraries of lead-like compounds. Our resulting battery of functional group filters is comprehensive, but hardly controversial and few if any attractive screening compounds are removed; indeed, we even retain a small number of compounds that might be regarded as controversial as we have discussed. To support our case, we fully disclose all our functional group definitions in the SI. We also issue a call to big pharma to release historical data to help further make this process as objective as possible.⁶¹

Exclusion of fingerprint-derived similars is a commonly accepted practice to select for new chemistries in preference to existing chemistries, the anticipation being that more biologically relevant space will consequently be covered. It is also essential for endeavors such as ours in order to reduce the size of excessively populated compound classes that may be present in vendor databases as we have shown. Such processes can be opaque to medicinal chemists, and this could be an impediment to drug discovery. Therefore, we tracked a particular compound class through such similarity filtering to illuminate for the medicinal chemist what effect this has on a potential SAR set. This exercise was also crucial in illustrating why it is not sensible to screen just the Stage 6 library in the absence of other libraries on which it was predicated, in particular the Stage 4 and 5 libraries. We also illustrated with a real SAR set why we believe that better analogue representation for poorly represented classes in vendor libraries combined with higher primary screening concentrations than often used could increase the odds of hit discovery and that this is critically important for difficult targets. As well as providing insights into the nature of vendor analogue sets, our exercises in similarity filtering highlights the limitations of retrieving the best analogue sets for SAR elaboration using such means. This is important because increasingly widespread access to HTS means that in many cases identification of screening hit analogues is undertaken by academic biologists using similarity rather than substructure searching of vendor sites because the intricacies of the latter are outside their knowledge base.

For the first time we also detail the use of our PAINS filters in establishing a major library expansion and show that this is associated with the exclusion of some 162,000 compounds. Also for the first time, we have deliberately titrated a PAINS-rejected set of screening hits to illustrate the subversive nature of these compounds and why we can justify their exclusion in library expansions.

Our libraries are accessible under certain conditions, and interested parties are encouraged to contact the author for more information. We have listed a random selection of 1000 compounds from the final stages of our Stage 6 compound selection and suggest they represent a grouping of the highest quality library of vendor-supplied lead-like compounds that are publicly accessible. By implication, vendor libraries bought with little or no triage are likely to comprise only a small portion of compounds that meet robust lead-like criteria.

The question then arises as to how we could now achieve a new library expansion of 400,000 compounds were we to undertake such a task in the immediate future. The evidence we have presented here is that this would not be possible without relaxation of our filters. We could therefore allow in some currently excluded groups or physicochemically less leadlike compounds (higher mw, higher cLogP, etc.) though in both cases the net effect would be importation of related groupings of generally less attractive compounds. We could also relax our similarity filters, and we have already discussed how this may be useful, although the main result would simply be increased analogue representation.

Of course, a preferred situation would be if lead-like screening space was more populated with diverse chemistries and library expansion could take place this way. Currently this is not possible, and the dire need for more lead-oriented synthesis has recently been discussed.²⁰ Indeed, it appears increasingly clear that vendor space is represented by some highly populated compound classes with much underpopulated or unpopulated space.⁶² This is likely to reflect relative synthetic ease. Given the need to better fill lead-like space, we can see no reason why any vendor effort should be spent on making compounds with a molecular weight greater than 400. These comments are not intended as a criticism. Vendors these days are generally highly professional outfits reliably supplying screening compounds of high quality, and it may not be cost competitive to explore more difficult chemistries. Many vendors are aware of the need for new lead-like chemistries and are actively addressing this issue. In the face of overwhelming chemical space to fill, sophisticated chemoinformatics methods^{63,64} can be used to prioritize compounds that fill chemical diversity holes. Until such time as this effort gains momentum, greater accessibility to corporate libraries has been a recent and welcome initiative.⁶⁵

■ ASSOCIATED CONTENT

■ Supporting Information

In silico profiling of the Stage 6 library, structures of 1000 random compounds, and functional group filters are detailed. Also included is a summary of file processing command lines, definitions used for functional group counts in final library, details of late-stage processing for final compound selection, and discussion on depth of available lead-like chemistry space. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +61 3 9903 9044. E-mail: Jonathan.Baell@monash.edu. Corresponding author address: Monash Institute of Pharmaceutical Sciences, Monash University (Parkville Campus), Parkville, Victoria, Australia, 3052.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the following: the financial support of the NHMRC (IRISS grant number 361646; also a Senior Research Fellowship for the author) and Victorian Stage Government OIS grant; Hendrik Falk for undertaking the PAINS titrations; John Parisot for organizing final purchase and shipping of Stage 6 compounds; and Phil Cruz and Tom Jones (Certara) for helpful software support.

■ ABBREVIATIONS:

HAT, histone acetyl transferase; HBA, hydrogen bond acceptor(s); HBD, hydrogen bond donor(s); HTS, high throughput screening; PAINS, pan-assay interference compounds; PK, pharmacokinetic; Ro3, rule-of-three; SAR, structure–activity relationships; SLN, sybyl line notation

■ 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 Discovery* **2011**, *10*, 188–195.
- (2) Frearson, J. A.; Collie, I. T. HTS and hit finding in academia - from chemical genomics to drug discovery. *Drug Discovery Today* **2009**, *14*, 1150–1158.
- (3) 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.
- (4) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470–479.
- (5) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902.
- (6) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382–384.
- (7) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug Delivery Rev.* **2002**, *54*, 255–271.
- (8) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH molecular libraries initiative. *Science* **2004**, *306*, 1138–1139.
- (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) Extensive discussion about one of the most prominent classes of PAINS, alkylidene rhodanines, may be found on Derek Lowe's well-known blog "In the pipeline". The prevailing view is overwhelmingly against such compounds being useful.
- (11) Tomasic, T.; Masic, L. P. Rhodanine as a scaffold in drug discovery: a critical review of its biological activities and mechanisms of target modulation. *Expert Opin. Drug Discovery* **2012**, *7*, 549–560.
- (12) Mendgen, T.; Steuer, C.; Klein, C. D. Privileged scaffolds or promiscuous binders: a comparative study on rhodanines and related heterocycles in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 743–753.

- (13) 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.
- (14) Dandapani, S.; Marcaurelle, L. A. Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.* **2010**, *6*, 861–863.
- (15) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877.
- (16) This number arises from the summation of the Stage 2 and 3 libraries (29K), Stage 4 (15K), Stage 5 (136K) and Stage 6 (93K primary screening set and excluding secondary set) libraries, multiplied by 100/80 to account for the 20% of vendor chemistry space unsolicited. Several thousand compounds were excluded from the 3rd party Stage 2 and 3 libraries because of poor lead-like properties, leading to their total of 29K rather than 37K compounds.
- (17) Verheij, H. J. Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Diversity* **2006**, *10*, 377–388.
- (18) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (19) 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.
- (20) Nadin, A.; Hattotuwegama, C.; Churcher, I. Lead-oriented synthesis: a new opportunity for synthetic chemistry. *Angew. Chem., Int. Ed.* **2012**, *51*, 1114–1122.
- (21) Voigt, J. H.; Bienfait, B.; Wang, S. M.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702–712.
- (22) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
- (23) Reymond, J. L.; Blum, L. C.; van Deursen, R. Exploring the chemical space of known and unknown organic small molecules at www.gdb.unibe.ch. *Chimia* **2011**, *65*, 863–867.
- (24) 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* (11), 2864–2875.
- (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) Renner, S.; Popov, M.; Schuffenhauer, A.; Roth, H. J.; Breitenstein, W.; Marzinzik, A.; Lewis, I.; Krastel, P.; Nigsch, F.; Jenkins, J.; Jacoby, E. Recent trends and observations in the design of high-quality screening collections. *Future Med. Chem.* **2011**, *3*, 751–766.
- (27) Yu, B.; Reynisson, J. Bond stability of the "undesirable" heteroatom-heteroatom molecular moieties for high-throughput screening libraries. *Eur. J. Med. Chem.* **2011**, *46*, 5833–5837.
- (28) Kalgutkar, A. S.; Gardner, I.; Obach, R. S.; Shaffer, C. L.; Callegari, E.; Henne, K. R.; Mutlib, A. E.; Dalvie, D. K.; Lee, J. S.; Nakai, Y.; O'Donnell, J. P.; Boer, J.; Harriman, S. P. A comprehensive listing of bioactivation pathways of organic functional groups. *Curr. Drug Metab.* **2005**, *6*, 161–225.
- (29) Huth, J. R.; Song, D.; Mendoza, R. R.; Black-Schaefer, C. L.; Mack, J. C.; Dorwin, S. A.; Lador, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Hajduk, P. J. Toxicological evaluation of thiol-reactive compounds identified using a La assay to detect reactive molecules by nuclear magnetic resonance. *Chem. Res. Toxicol.* **2007**, *20*, 1752–1759.
- (30) Snodin, D. J. Genotoxic impurities: from structural alerts to qualification. *Org. Process Res. Dev.* **2010**, *14*, 960–976.
- (31) Edwards, P. J.; Sturino, C. Managing the liabilities arising from structural alerts: a safe philosophy for medicinal chemists. *Curr. Med. Chem.* **2011**, *18*, 3116–3135.
- (32) Smith, G. F. Designing drugs to avoid toxicity. *Prog. Med. Chem.* **2011**, *50*, 1–47.
- (33) 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 NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* **2005**, *127*, 217–224.
- (34) Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 139–144.
- (35) Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.
- (36) Baell, J. B. Observations on screening-based research and some concerning trends in the literature. *Future Med. Chem.* **2010**, *2*, 1529–1546.
- (37) Baell, J. B. Redox-active nuisance screening compounds and their classification. *Drug Discovery Today* **2011**, *16*, 840–841.
- (38) Huggins, D. J.; Venkitaraman, A. R.; Spring, D. R. Rational methods for the selection of diverse screening compounds. *ACS Chem. Biol.* **2011**, *6*, 208–217.
- (39) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* **2010**, *53*, 37–51.
- (40) Hack, M. D.; Rassokhin, D. N.; Buyck, C.; Seierstad, M.; Skalkin, A.; ten Holte, P.; Jones, T. K.; Mirzadegan, T.; Agrafiotis, D. K. Library enhancement through the wisdom of crowds. *J. Chem. Inf. Model.* **2011**, *51*, 3275–3286.
- (41) Babaoglu, K.; Simeonov, A.; Lrwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *J. Med. Chem.* **2008**, *51*, 2502–2511.
- (42) Groll, M.; Kim, K. B.; Kairies, N.; Huber, R.; Crews, C. M. Crystal structure of epoxomicin: 20S proteasome reveals a molecular basis for selectivity of alpha 'beta'-epoxyketone proteasome inhibitors. *J. Am. Chem. Soc.* **2000**, *122*, 1237–1238.
- (43) Tse, C.; Shoemaker, A. R.; Adickes, J.; Anderson, M. G.; Chen, J.; Jin, S.; Johnson, E. F.; Marsh, K. C.; Mitten, M. J.; Nimmer, P.; Roberts, L.; Tahir, S. K.; Mao, Y.; Yang, X. F.; Zhang, H. C.; Fesik, S.; Rosenberg, S. H.; Elmore, S. W. ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. *Cancer Res.* **2008**, *68*, 3421–3428.
- (44) Baker, S. J.; Ding, C. Z.; Akama, T.; Zhang, Y. K.; Hernandez, V.; Xia, Y. Therapeutic potential of boron-containing compounds. *Future Med. Chem.* **2009**, *1*, 1275–1288.
- (45) Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **2011**, *54*, 2529–2591.
- (46) Oballa, R. M.; Truchon, J. F.; Bayly, C. I.; Chaudet, N.; Day, S.; Crane, S.; Berthelette, C. A generally applicable method for assessing the electrophilicity and reactivity of diverse nitrile-containing compounds. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 998–1002.
- (47) Janetka, J. W.; Almeida, L.; Ashwell, S.; Brassil, P. J.; Daly, K.; Deng, C.; Gero, T.; Glynn, R. E.; Horn, C. L.; Ioannidis, S.; Lyne, P.; Newcombe, N. J.; Oza, V. B.; Pass, M.; Springer, S. K.; Su, M.; Toader, D.; Vasbinder, M. M.; Yu, D.; Yu, Y.; Zabudoff, S. D. Discovery of a novel class of 2-ureido thiophene carboxamide checkpoint kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4242–8.
- (48) Zabudoff, S. D.; Deng, C.; Grondine, M. R.; Sheehy, A. M.; Ashwell, S.; Caleb, B. L.; Green, S.; Haye, H. R.; Horn, C. L.; Janetka, J. W.; Liu, D.; Mouchet, E.; Ready, S.; Rosenthal, J. L.; Queva, C.; Schwartz, G. K.; Taylor, K. J.; Tse, A. N.; Walker, G. E.; White, A. M. AZD7762, a novel checkpoint kinase inhibitor, drives checkpoint abrogation and potentiates DNA-targeted therapies. *Mol. Cancer Ther.* **2008**, *7*, 2955–66.
- (49) Reddick, J. J.; Cheng, J. M.; Roush, W. R. Relative rates of Michael reactions of 2'-(phenethyl)thiol with vinyl sulfones, vinyl sulfonate esters, and vinyl sulfonamides relevant to vinyl sulfonyl cysteine protease inhibitors. *Org. Lett.* **2003**, *5*, 1967–1970.

(50) Palmer, J. T.; Rasnick, D.; Klaus, J. L.; Bromme, D. Vinyl sulfones as mechanism-based cysteine protease inhibitors. *J. Med. Chem.* **1995**, *38*, 3193–3196.

(51) Falk, H.; Connor, T.; Yang, H.; Loft, K. J.; Alcindor, J. L.; Nikolakopoulos, G.; Surjadi, R. N.; Bentley, J. D.; Hattarki, M. K.; Dolezal, O.; Murphy, J. M.; Monahan, B. J.; Peat, T. S.; Thomas, T.; Baell, J. B.; Parisot, J. P.; Street, I. P. An efficient high-throughput screening method for MYST family acetyltransferases, a new class of epigenetic drug targets. *J. Biomol. Screening* **2011**, *16*, 1196–1205.

(52) Lu, Q.; Quinn, A. M.; Patel, M. P.; Semus, S. F.; Graves, A. P.; Bandyopadhyay, D.; Pope, A. J.; Thrall, S. H. Perspectives on the discovery of small-molecule modulators for epigenetic processes. *J. Biomol. Screening* **2012**, *17*, 555–571.

(53) Arrowsmith, C. H.; Bountra, C.; Fish, P. V.; Lee, K.; Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discovery* **2012**, *11*, 384–400.

(54) Furdas, S. D.; Kannan, S.; Sippl, W.; Jung, M. Small molecule inhibitors of histone acetyltransferases as epigenetic tools and drug candidates. *Arch. Pharm. (Weinheim, Ger.)* **2012**, *345*, 7–21.

(55) Rye, C. S.; Baell, J. B. Phosphate isosteres in medicinal chemistry. *Curr. Med. Chem.* **2005**, *12*, 3127–3141.

(56) Bista, M.; Smithson, D.; Pecak, A.; Salinas, G.; Pustelny, K.; Min, J.; Pirog, A.; Finch, K.; Zdzalik, M.; Waddell, B.; Wladyka, B.; Kedracka-Krok, S.; Dyer, M. A.; Dubin, G.; Guy, R. K. On the mechanism of action of SJ-172550 in inhibiting the interaction of MDM4 and p53. *PLoS One* [Online] **2012**, *7*, Article 6. <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0037518> (accessed Aug 10, 2012).

(57) Calamini, B.; Silva, M. C.; Madoux, F.; Hutt, D. M.; Khanna, S.; Chalfant, M. A.; Saldanha, S. A.; Hodder, P.; Tait, B. D.; Garza, D.; Balch, W. E.; Morimoto, R. I. Small-molecule proteostasis regulators for protein conformational diseases. *Nat. Chem. Biol.* **2012**, *8*, 185–96.

(58) van Delft, M. F.; Wei, A. H.; Mason, K. D.; Vandenberg, C. J.; Chen, L.; Czabotar, P. E.; Willis, S. N.; Scott, C. L.; Day, C. L.; Cory, S.; Adams, J. M.; Roberts, A. W.; Huang, D. C. S. The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. *Cancer Cell* **2006**, *10*, 389–399.

(59) Lagorce, D.; Maupetit, J.; Baell, J.; Sperandio, O.; Tuffery, P.; Miteva, M. A.; Galons, H.; Villoutreix, B. O. The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound collections. *Bioinformatics* **2011**, *27*, 2018–2020.

(60) 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.

(61) While this manuscript was under review, Lilly Research Laboratories published a compilation of disfavored compounds for drug development.⁶⁶

(62) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Mod.* **2011**, *51*, 2174–2185.

(63) Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325–330.

(64) Reymond, J. L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30–38.

(65) For an example where corporate libraries can be accessed under appropriate arrangements see www.lcgcinc.com (accessed 20th Nov 2012).

(66) Bruns, R. F.; Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **2012**, *55* (22), 9763–9772.