Rational Methods for the Selection of Diverse Screening Compounds

David J. Huggins,[†,‡,§] Ashok R. Venkitaraman,[‡] and David R. Spring[‡,§,*]

[†]TCM Group, Cavendish Laboratory, University of Cambridge, 19 J J Thomson Avenue, Cambridge CB3 0HE, United Kingdom

[‡]Cambridge Molecular Therapeutics Programme, The Medical Research Council Cancer Cell Unit, Hutchison/MRC Research Centre, University of Cambridge, Hills Road, Cambridge CB2 2XZ, United Kingdom

[§]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

ABSTRACT: Traditionally a pursuit of large pharmaceutical companies, high-throughput screening assays are becoming increasingly common within academic and government laboratories. This shift has been instrumental in enabling projects that have not been commercially viable, such as chemical probe discovery and screening against high-risk targets. Once an assay has been prepared and validated, it must be fed with screening compounds. Crafting a successful collection of small molecules for screening poses a significant challenge. An optimized collection will minimize false positives while maximizing hit rates of compounds that are amenable to lead generation and optimization. Without due consideration of the relevant protein targets and the downstream screening assays, compound filtering and selection can fail to explore the great extent of chemical diversity and eschew valuable novelty. Herein, we discuss the different factors to be considered and methods that may be employed when assembling a structurally diverse compound collection for screening. Rational methods for selecting diverse chemical libraries are essential for their effective use in high-throughput screens.
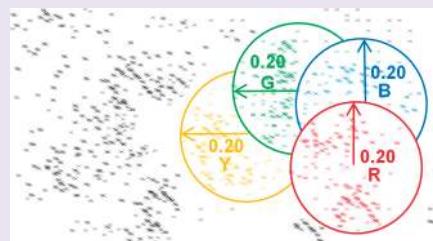
The earliest efforts in drug discovery focused on crude extracts from natural sources, and success relied mainly on trial and error. Work in the middle of the last century established the concept of a molecular disease,[1] moving drug discovery in a more rational direction and toward screening compounds against a molecular target. Natural products provided the majority of early drugs and still remain as an invaluable source of chemicals for screening, along with semisynthetic derivatives.[2] In more recent times, the advent of combinatorial chemistry provided a radical increase in the number of available screening compounds, and this was coupled with high-throughput screening (HTS) of large chemical libraries.[3] Despite many failures among the successes, HTS remains a widely used method for initiating the process of drug and chemical probe discovery.[4−9] The concept of a drug-like molecule has existed for many years[10] and includes optimized parameters for physicochemical properties as well as functional groups to be avoided. This concept has been extended to consider lead-like instead of drug-like molecules,[11] and this progresses naturally to the identification of hit-like molecules, which are geared to provide true positive results in HTS assays and yield a basis for lead generation.[12] The vastness of chemical space means that there are currently tens of millions of molecules available for purchase and screening. Even using harsh filters to remove unwanted compounds, there are on the order of a million hit-like molecules available commercially.[13,14] However, identifying a representative subset of these molecules to screen is a complex task, with multiple scientific, financial and logistical considerations. While this Review is unable to comprehensively cover the multifold aspects of library design, its aim is to highlight the key issues that must be taken into account. This is now important in academic groups and government laboratories as well as in industry.[15] Here we review current methods for crafting screening compound collections and outline the traps and pitfalls. This will be done in three sections: compound sourcing, compound filtering, and compound selection. Finally, we highlight key challenges to the field and outline future directions.

COMPOUND SOURCING

There are many suppliers of screening compounds, ranging from small chemical suppliers with hundreds of compounds to large ones with over a million compounds. Many collections of small molecules have been analyzed for drug-like and lead-like properties,[13,14,16−18,19] and chemical supplier libraries are being increasingly tailored toward these parameters. Details of the main screening libraries from six chemical suppliers with varied collections of over 300,000 screening compounds are reported in Table 1. At present, all have a high pass rate for commonly employed drug-like and lead-like filters. However, compound collections turn over rapidly and should be analyzed in this way prior to selecting suppliers. Compound prices per milligram vary widely dependent on the number of compounds purchased and the sample weight per compound required, with significantly lower prices per compound if thousands or tens of thousands are purchased. Theoretically, searching the entirety of currently available chemical space encompasses the maximum commercially available molecular diversity. In practice, a great expanse of available diversity can be sampled by selecting large numbers of

Table 1. Details of the Screening Libraries for Six Chemical Suppliers, the ZINC Database of Purchasable Molecules, and the Drugbank Database of Experimental Drugs[a]

compound source	compound collection	url	no. of compounds	% Lipinski passes	% REOS passes
Asinex	Gold and Platinum Collections	http://www.asinex.com	364,407	79.6	73.0
Chembridge	Express Pick Library	http://www.chembridge.com	442,051	84.0	66.6
ChemDiv	Discovery Chemistry	http://www.chemdiv.com	789,603	73.8	72.1
Enamine	HTS Collection	http://www.enamine.net	1,116,406	90.7	79.6
Life Chemicals	Stock	http://www.lifechemicals.com	327,211	84.9	76.6
Vitas M Laboratories	HTS Stock	http://www.vitasmlab.com	476,184	75.1	65.8
Drugbank	All Drugs	http://www.drugbank.ca	4,886	71.4	51.7
ZINC	Purchasable Compounds	http://zinc.docking.org	18,671,085	87.2	73.1

[a] All physicochemical properties were generated with Qikprop, and filtering was performed with Canvas. The compound collection refers to the subset of molecules that was analyzed from each source.

compounds from a few chemical suppliers with diverse collections. Many chemical suppliers also sell preselected diverse libraries at reduced cost. These are generally selected by rational means, but the compound filters employed may have been too harsh or too lenient, dependent on the nature of the screening assay and the target. Furthemore, although the compounds tend to be relatively diverse, they are also much more likely to have been tested by other laboratories, as they are for sale off-the-shelf. Including novelty in HTS is a vital aspect of drug discovery, and many firms offer unlisted libraries at higher costs, promising an easier path to intellectual property rights.

Compound Databases. In addition to compound libraries direct from chemical suppliers, there are a number of preassembled online data repositories including ZINC[20] (http://zinc.docking.org/), emolecules (http://www.emolecules.com/), and Chemspider (http://www.chemspider.com/). The ZINC repository currently has the largest number of compounds, including the complete compound libraries of the majority of chemical suppliers. The number of molecules in the ZINC set of purchasable compounds currently stands at just under 18.7 million. However, chemical suppliers commonly update their libraries every few months, which may not be reflected in data repositories such as ZINC. Despite the huge number of commercially available compounds, existing chemistry efforts have probed only a small proportion of chemical space. The number of synthetically feasible, drug-like molecules is estimated to be in excess of 10^{60},[21] and only a small subset of this has been explored. For example, data compiled in the Generated Database of Molecules (http://www.dcb-server.unibe.ch/groups/reymond/gdb/start.html) demonstrate that less than 0.5% of the synthetically feasible compounds comprised of up to 11 atoms of C, N, O, and F are recorded in public databases as having been synthesized.[22] Recent studies have also highlighted a large number of novel ring systems that are not currently represented in available chemical space.[23] Many sources of diversity are excluded from existing compound collections, and this greatly restricts the coverage of chemical space. In particular, the bias against chirality skews commercially available compounds toward flat compounds with many aromatic rings.[24] This in turn may negatively impact on the properties related to absorption, distribution, metabolism, elimination, and toxicity (ADMET) and increase the risk of attrition during development.[25] Shelat and Guy have questioned whether libraries of synthetic molecules are suitable for addressing novel drug targets and suggest the use of natural products in HTS, particularly for phenotypic and high-content screens.

Natural Products. The vast majority of commercially available small molecules are obtained from synthetic chemistry. Nonetheless, nature is an important source of biologically active compounds, and natural products have played a key role in drug discovery efforts. It has been estimated that as many as 50% of marketed small molecule drugs have been derived from natural products.[26] However, of the compounds currently approved for marketing each year, natural products represent a much lower percentage. Many chemical suppliers sell natural products for HTS, and some chemical suppliers specialize in natural product chemistry. The natural product collections are usually separated from synthetic compounds and can be significantly more expensive. However, they can provide unique chemical structures and may show more drug-like ADMET properties.[27] Natural products have proven particularly powerful as anticancer and anti-infective agents[2] and tend to be well suited to phenotypic screening. Recent analysis shows that there are many ring systems present in natural products that are not found in screening libraries, and many have suggested that screening compounds should be further biased toward biogenic scaffolds.[28,29] However, the advantages of natural products must be balanced against their often greater structural complexity that may lead to difficulties in synthesis and purification of analogues during lead generation and optimization. There is still great controversy over the relative merits of screening natural products or natural product derivatives versus screening libraries from combinatorial chemistry or diversity oriented synthesis.[30] Both have advantages and disadvantages, and thus HTS libraries commonly combine both sources, though typically with more synthetic small molecules. Recently, it has been suggested that compounds balancing the properties of natural products and synthetic molecules may be optimal.[31]

In summary, there are multiple sources of potential screening compounds, and successful libraries typically strike a balance between synthetic compounds and natural products. However, although the growth in commercially available chemical space should always be capitalized upon, many compounds are unsuitable for screening in HTS assays and should be filtered out of any quality screening collection.

■ COMPOUND FILTERING

In order to obtain commercially available hit-like compounds, computational filters are commonly used to remove compounds with undesirable properties. Ideal drug-like and lead-like molecules have differing properties, and these differ again from hit-like molecules. In general, the physicochemical properties of a lead-like
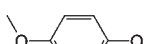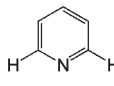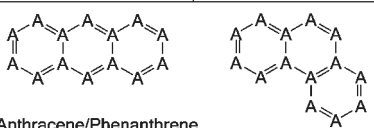
Figure 1. Chemical structures used in compound filtering. Chemical structures of functional groups commonly used to remove compounds from consideration in HTS assays. The functional group name and SMILES/SMARTS string used in the filter are reported.

molecule can be improved during lead optimization toward a drug-like molecule by tailoring the lipophilicity. Similarly, the binding affinity of a hit-like molecule can be improved during the process of hit explosion to yield a lead-like molecule. However, hit-like molecules must be large and lipophilic enough to gain sufficient binding affinity that they can be identified in a screening assay, but not so large that they have a very small probability of binding. Larger and more complex molecules have a lower probability of exhibiting perfect shape and electrostatic complementarity with any given target, and this suggests that smaller and less complex molecules will more commonly provide starting points for drug development.[32] An ideal hit molecule should also be amenable to chemical elaboration, show reasonable levels of cell permeability, and have a range of commercially available analogues, some of which have also been tested in the same assay.

Computational Filters. There are numerous computational filters used to mark compounds that may have problems due to assay interference or downstream ADMET properties. The most commonly used of these are physicochemical property filters that specifically attempt to remove compounds that may lead to low levels of drug absorption and distribution. An exception that is ignored by these filters is compounds that are substrates for drug transporters, which recent work suggests may be a significant proportion of molecules.[33] In addition to Lipinski's well-known rule of five,[34] Ghose filters[35] and Veber filters[36] are commonly employed to filter compounds. Noteworthy analysis has also been performed by Walters,[37] Oprea,[38] Egan,[39] Lee,[40] Baurin,[13] and Martin.[41] The key properties that determine drug absorption and distribution for an oral drug are the lipophilicity measures of the octanol/water partition coefficient ($\log P$) and surface area of the polar atoms in the molecule (PSA).[42−44] Analysis of trends in launched drugs has highlighted a significant increase in molecular weight in the past 50 years, but a negligible increase in $\log P$ values.[45] This is not surprising, as drugs with increased $\log P$ tend to be more promiscuous binders and can thus be expected to have a higher attrition rate in later development.[46] However, studying the most recent trends in molecules being synthesized in leading drug discovery companies suggests an increase in both molecular weight and $\log P$.[45] This has been attributed to the fact that more lipophilic drugs have the potential to be more efficacious, as they tend to have increased binding affinity. It has been suggested that this may adversely affect drug attrition rates in the

future due to an increased likelihood of toxicity.[47] However, as discussed, larger and more complex molecules have a lower probability of exhibiting perfect shape and electrostatic complementarity with any given target and are thus expected to show greater specificity.[32] This predicted increase in promiscuity due to increased lipophilicity may thus be ameliorated by increased complexity. Despite the noted increase in molecular weight, there is great pressure during the development process to lower the molecular weight, likely because larger molecules show reduced passive absorption across cell membranes, increased number of toxic pharmacophores, or rapidly metabolized moieties.[48] One caveat when filtering on lipophilicity or solubility is to note whether experimental values or predicted values are being used. Solubility predictions based on clog P values or PSA can be accurate in some circumstances but are inaccurate in others and tend to perform particularly badly for charged compounds.[49] Charged compounds may be better represented by the octanol/water distribution coefficient log D, which takes into account the different protonation states. It is vital to carefully consider whether compounds should be excluded on the basis of predicted insolubility, when such predictions can be inaccurate.

One other significant method for marking ADMET risks are the rapid elimination of swill[50] (REOS) filters. As well as physicochemical properties, REOS filters remove molecules containing certain functional groups, as described by SMILES or SMARTS patterns.[51] Some of these are shown in Figure 1. REOS filters flag compounds containing functional groups that may lead to false positives due to reactivity or assay interference, which have long been noted as a problem in HTS efforts.[52] They also remove compounds containing functional groups known to be risks for ADMET. However, it is important to note that many known drug molecules fail the common physicochemical and substructure filters. The Drugbank[53] (http://www.drugbank.ca/) contains structural data for over 1,350 FDA-approved small molecule drugs and nearly 5000 experimental drug entries. Analysis of the Drugbank experimental drugs is shown in Table 1 and reveals that only 71.4% pass all of the Lipinski filters and only 51.7% pass all of the REOS substructure filters. This data highlights that compound filtering is used to reduce risk but will also eliminate useful molecules from further consideration. More recently, a Herculean analysis of compounds hitting multiple orthogonal HTS assays has led to the identification of pan assay

Table 2. Details of Physicochemical Property Filters To Mark Drug-like and Lead-like Compounds for Screening Libraries

	MW	PSA (Å^2)	HBA	HBD	log P	rotatable bonds	no. of atoms	charge
Lipinski (1997)	≤500		0 to 10	0 to 5	≤5.0			
Ghose (1999)	160 to 480				−0.4 to +5.6		20 to 70	
Oprea Drug-Like (2000)			2 to 9	0 to 2		2 to 8		
Egan (2000)		≤130			−1.0 to +5.8			
Walters (2000)	200 to 500	≤120	0 to 10	0 to 5		0 to 8	20 to 70	−2 to +2
Oprea Lead-Like (2001)	≤450		0 to 8	0 to 5	−3.5 to +4.5			
Veber (2002)		≤140				0 to 10		
REOS (2002)	200 to 500		0 to 10	0 to 5	−5.0 to +5.0	0 to 8		−2 to +2
Martin (2005)		≤150						

interference compounds (PAINS).[54] As increasing amounts of assay data from different HTS efforts around the world is becoming publically available, a clearer picture of compounds and functional groups that tend to yield false positives is developing.[55] This development is vital, as frequent hitters are likely to be over represented in compounds from chemical vendors due to an increased likelihood that they will be ordered as analogues of apparent hits. Research has also specifically highlighted substructures that alert when a compound may be a DNA-reactive genotoxin.[56] While this may be acceptable in a screening hit, it would almost certainly have to be removed in the hit to lead process.

Physicochemical Property Filters. The majority of physicochemical property filters are simple to understand. Eight drug-like filters and one lead-like filter are described in Table 2. There is general agreement, although the exact properties vary slightly. Any of these rules can be used, alone or in conjunction, to filter a set of compounds, and it is worth noting that many of the properties are highly correlated, such as log P and PSA. However, due consideration must be given to the details of the screening assay and the nature of the target as this affects the desired physicochemical properties of the screening compounds. For example, a fragment with a molecular weight of 200 may be too small to show measurable binding in typical HTS assays or compete with high-affinity ligands. However, if the assay is tailored to identify smaller molecules, fragment-based methods have been shown to be very useful, with higher ligand efficiencies[57] and a greater potential for chemical elaboration and linking.[58] Compound filters for fragments are completely different from filters for traditional small molecules. Phenotypic screens also place a different pressure on the screening library, with considerably more emphasis on cell permeability at the initial stage. As well as the importance of the assay format, the composition of an ideal screening library also varies with the protein target. Many existing screening libraries and are tailored toward screening against a narrow range of targets such as kinases and GPCRs.[59] A screening library tailored toward screening against protein−protein interactions would have a very different profile. Recent analysis collected in the TIMBAL database[23] suggests that inhibitors of protein−protein interactions have higher molecular weights and lipophilicity than inhibitors of buried binding sites, as well as a greater number of hydrogen bond donors, hydrogen bond acceptors, and rotatable bonds. While the general applicability of this approach to generating approved drugs remains to be seen, it is an important consideration. As well as traditional physicochemical property filters, there are now a number of flags for more complex properties.[60] Increasing evidence shows that small molecules may cause nonspecific protein aggregation[61] and

thus lead to false positives in some assays. Experimental work has shown that a significant number of compounds may act in this way and potential risks can be identified and removed from consideration.[62] There are also experimental methods to identify compounds that are reactive, such as ALARM NMR,[63] and also for compounds containing fluorophores.[64] However, while the latter is of great importance for fluorometric assays, it is of little or no importance in other assays. Experimental studies such as PAINS have identified molecular scaffolds that form the basis for promiscuous inhibitors and thus yield false positives in many screening assays.[54,65] Defining the mechanism underlying the promiscuous inhibition of these PAINS compounds will no doubt provide significant but interesting challenges in the next decade. In addition there are now methods for predicting compounds that disrupt particular screening assays,[66] but these methods are approximate and should be used with this understanding.

Substructure Filters. Many filters simply remove compounds with specific functional groups that are known to interfere with HTS assays or cause problems later in drug development. The importance of removing these functional groups has been discussed in numerous papers.[37,52] The majority of screening libraries contain very few if any of the most troublesome compounds such as aldehydes, epoxides, or α-halo ketones. The prevalence of these three groups in the six supplier databases is on average 0.3%, 0.01%, and 0.04%, respectively. However, many still contain potential risks such as isolated alkenes (12.3%), αβ-unsaturated carbonyls (8.5%), or nitro groups (7.6%). The prevalence of the more common functional groups can be seen in Table 3. Each of these substructures is a potential liability for the following reasons:

- 1,2-dicarbonyl: metabolically unstable/potential toxicity due to mutagenicity
- 1,2-dimethoxy: prone to oxidation yielding reactive quinones.
- 1,4-dimethoxy: very prone to oxidation yielding reactive quinones
- αβ-unsaturated carbonyl: prone to reactivity by acting as a Michael acceptor
- acetal: metabolically unstable due to acetal hydrolysis
- acylhydrazide: metabolically unstable due to acyl hydrolysis
- aliphatic ketone: metabolically unstable due to nucleophilic attack
- alkene: metabolically unstable due to epoxidation
- aminothiazole: potential toxicity
- anthracene/phenanthrene-like: known DNA intercalation
- nitro group: prone to reduction yielding reactive species/potential hepatocarcinogens

Table 3. Percentage of Compounds Failing Common Drug-Like Filters for Unfavorable Physiochemical Properties and Unwanted Substructures for the Six Combined Chemical Supplier Libraries, the ZINC Database of Purchasable Molecules and the Drugbank Database of Experimental Drugs[a]

	combined suppliers	Drugbank	ZINC
clog $P > 5$	15.8	7.0	10.7
HBA > 10	3.8	23.0	6.7
HBD > 5	0.0	13.1	0.1
MW > 500	4.9	13.3	1.7
PSA > 150	1.8	22.0	3.3
rotatable bonds >10	1.5	20.3	2.5
isolated alkene	9.1	12.3	8.7
$\alpha\beta$-unsaturated carbonyl	8.5	8.5	6.9
1,2-dimethoxy	7.6	6.0	7.6
nitro	7.4	6.6	6.5
acylhydrazide	4.0	4.6	4.1
aminothiazole	4.0	4.8	3.1
thiourea	3.3	4.3	1.6
anthracene/phenanthrene-like	3.1	5.9	1.2
unflanked pyridyl	3.1	5.9	2.5
acetal	2.7	13.0	2.0
methylene-dioxy	2.3	4.6	1.5
aliphatic ketone	2.1	10.6	2.0
1,2 dicarbonyl	1.6	5.6	1.0
1,4-dimethoxy	1.5	4.5	1.6

[a] All physicochemical properties were generated with Qikprop and filtering was performed with Canvas.

- methylenedioxy: metabolically unstable due to acetal hydrolysis/prone to oxidation yielding reactive quinones
- thiourea: metabolically unstable due to flavin oxidation/potential onspecific protein binding
- unflanked pyridyl: potential interference with cytochrome P450s due to metal ion coordination

However, many of these functional groups do appear in certified drug molecules,[67] as shown in Table 3, and many show no activity in HTS assays.[68] When eliminating functional groups due to any ADMET risk, the nature of the functional group should be considered. It may be easier to replace a potentially risky side-group at the hit-to-lead stage than a potentially risky core group. For example, a nitroaromatic side-group can be replaced with another similar side-group such as a trifluoromethanesulfonyl side-group to retain or increase binding affinity without disrupting the structure of the molecule.[69] The same is not true for a 2-aminothiazole core group, as its shape and hydrogen bonding characteristics are more difficult to mimic without disrupting the structure of the molecule. Despite this, scaffold hopping can be achieved and is increasingly common.[70] When eliminating functional groups due to the risk of cytotoxicity, it is important to consider the target, as some therapies (for cancer in particular) are damaging to cells. For example, 2-aminothiazoles may lead to cytotoxicity, but they form the basis of a number of potent CDK inhibitors for cancer therapy.[71] Functional groups implicated in organ toxicity may also be acceptable in chemical probe discovery.

Filtering Tools. There are a number of software packages used to predict chemical properties and/or filter screening compounds. This includes Accelrys' Pipeline Pilot,[72] MOE's sdfilter,[73] Schrodinger's qikprop,[74] and Openeye's filter,[75] which is freely available

to academics. Once the filtering process is complete, it is important to inspect a subset of the resulting structures. No matter how sophisticated the filtering criteria and algorithms, a scientist should always ensure that the remaining compounds meet their requirements. Despite the importance of filtering compounds to prevent screening potentially problematic compounds, it is common to screen a small proportion of "wildcards" that do not pass all of the filters. As seen in Tables 1 and 2, many drug molecules do not pass the drug-like or lead-like filters and contain significant proportions of functional groups that are commonly removed by HTS filters. For example, the REOS rule to exclude compounds with more than four joined rings removes all steroids and nearly 10% of the Drugbank experimental drugs. It is important to realize that the process of compound filtering is about minimizing risk and downstream expenditure rather than maximizing hit rate. For example, reactive groups may present the risk of false positives, but work has shown that this is not always the case.[68] In some cases, reactive groups can act as covalent inhibitors, inactivating the target by binding irreversibly, and thus provide an advantage over noncovalent inhibitors. However, this activity may be difficult to extract from HTS data as it can be hard to discriminate from unwanted reactivity. Potentially reactive compounds should remain, at most, a small percentage of any screening library, unless there is a clear plan to extract useful data on covalent inhibition from the screening assay.

In summary, it may be necessary to rethink the process of designing libraries for screening against the more diverse range of targets now being considered. Research at Harvard,[76] the NIH,[6,77] and the DDU in Dundee[9] among others has shown that HTS is feasible in a nonindustrial center and can be vital in developing treatments for neglected diseases. While such drug development projects must also select screening compounds with care, many of the functional group and physicochemical property filters are unsuitable for screening efforts aimed at development of chemical probes. Compounds causing assay interference or low solubility should be avoided, but compounds causing liver toxicity or poor oral absorption may be acceptable. Recent analysis suggests that the nature of screening hits is shifting to larger and more lipophilic molecules as a result of the increased use of *in vitro* assays over *in vivo* assays.[78] This is expected to shift or widen the nature of screening libraries. However, the exact nature of the assay and the target must be considered when selecting compound exclusions as, for a diversity library aiming to span multiple assays and targets, it may not be appropriate to remove all potential risks. A balance must be reached between filtering out all compounds that are a risk in any drug development program and only filtering compounds that are a risk in all programs. There is now a critical mass of published data highlighting risks for compound interference, and this can easily be applied to hits post screening, along with experimental methods to detect false positives such as dose−response plotting. This should ensure that screening libraries take advantage of the enormous diversity in chemical space, while assessing risk appropriately. With respect to chemical diversity, chemical suppliers will only provide chiral compounds if there is a market for them, and thus filtering out chiral compounds from screening libraries will drive the purchasable chemical space further in this direction and away from biogenic chemical space.

■ COMPOUND SELECTION

Aggressive filtering may remove up to 50% of compounds from consideration, but huge numbers of commercially available
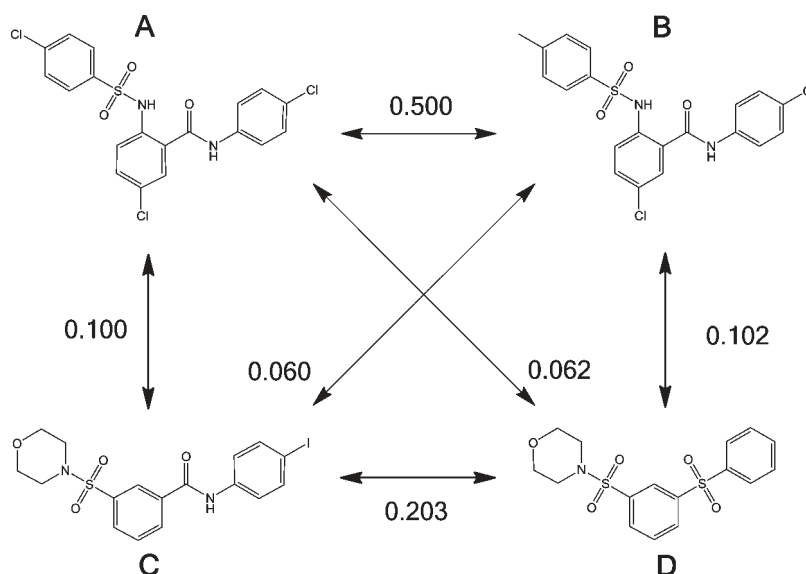
Figure 2. Example of similarity between compounds. Four compounds and the Tanimoto similarity between them. The compounds were assigned radial fingerprints using Schrodinger's Canvas software at 64-bit precision using daylight-invariant atom types.

compounds still remain. The main aim of compound selection is to pick a subset of these compounds for testing. In general, it is wasteful to test many compounds with similar structures in front-line assays, at the expense of more diverse compounds. Analysis has shown that if a compound is biologically active, a molecule with very high similarity will have a similar biologically activity, and thus testing the second molecule in the frontline assay is unlikely to be worthwhile.[79,80] It is thus common to select a structurally diverse subset of compounds that represents the chemical space being considered. However, chemical space grows very rapidly with molecular size, and in 200 years of chemical synthesis we have covered only a tiny fraction of chemical space up to a molecular weight of 500. The biggest screening libraries, which are of the order of tens of millions of molecules, can never hope to cover this space. Approaching compound selection in a sensible manner is thus very important.[81]

Measuring Chemical Diversity. Molecular similarity is a key prerequisite in assessing molecular diversity.[82] There are many different techniques to measure whether two compounds are similar,[80,83] but none of them are entirely satisfactory. From a pharmaceutical perspective, the ideal metric would predict that two compounds are similar if they elicit the same biological effect by hitting the same biological target and binding in the same pose. Unfortunately such a metric does not exist. Currently used metrics predict that two compounds are similar if they have similar chemical connectivity or similar shape and electrostatic form. One important issue in assessing chemical similarity is that a compound can be very different in its various conformations, tautomers, and protonation states. Two compounds that are calculated to be similar in specific tautomeric states may be calculated to be different in other states. However, there are numerous computational methods for the enumeration of protonation and tautomeric states. This includes Schrodinger's Ligprep,[84] the Openeye toolkit,[75] CCG's MOE,[73] Tripos Sybyl,[85] and Accelrys' Discovery Studio.[72] Three of the most common methods for predicting similarity are fingerprint,[86] shape-base,d[87] and pharmacophore[70] methods. These methods are commonly used in virtual screening when a known active compound has been identified. Fingerprint methods are relatively simple and usually

two-dimensional. Each molecule is assessed for a number of atom and bond connectivities. Each of these connected units is termed a bit/key, and the combination of bits/keys that are present in a given molecule is its fingerprint. Two molecules with similar fingerprints have similar atoms in similar bonding environments and are likely to bind in similar ways to a protein target. There are a number of fingerprinting techniques as well as a number of atom-typing schemes and close reading of the current literature is recommended before selecting a method, as this is still a rapidly developing field.[88] Recent analysis has shown that atom-type based radial fingerprints perform well,[89] but other work suggests that fingerprints based on physicochemical properties or pharmacophores may perform better.[90] Different fingerprinting methods can yield very different similarities, and thus an exact comparison with literature is not always appropriate. There are also a number of similarity/difference metrics,[91] and while the Tanimoto metric is most commonly used, close reading of the current literature is again recommended. The molecules in Figure 2 were analyzed using radial fingerprints based on daylight atom types using Schrodinger's Canvas software, and Tanimoto similarity scores were then generated. As can be seen, molecule with a high similarity such as A and B are very similar and would likely give similar assay results, whereas molecule A and D are significantly different and should ideally both be tested in a frontline assay. Shape-based methods compare molecules by analyzing whether they have the same shape and electronic form. This is implemented in Openeye's ROCS and EON software,[75] which is widely used and is freely available to noncommercial groups working toward public disclosure.[92] Pharmacophore methods have the obvious advantage of including the three-dimensional geometry of the molecules. As noted, chemical similarity is a very important concept in assessing chemical diversity. While three-dimensional methods have the potential to provide a much more accurate model of molecular similarity, there is great difficulty in applying them when the bioactive conformation is unknown, as is the case in diversity analysis. Thus, two-dimensional methods such as fingerprinting remain the tool of choice at present.

Rational Selection. Once a set of compounds has been analyzed on the basis of similarity, it is possible to select a diverse
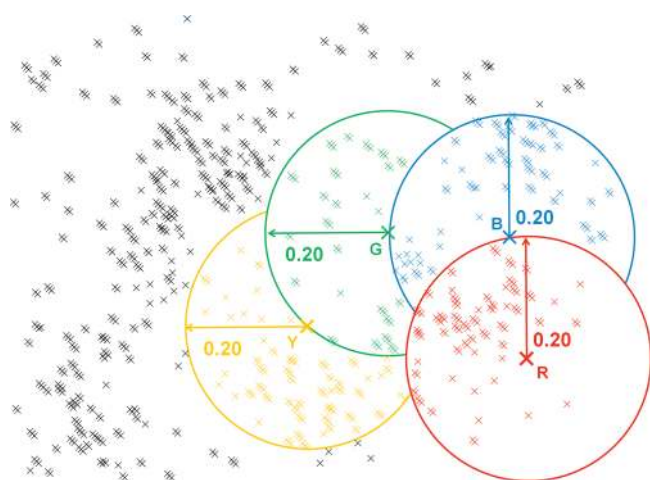
Figure 3. Clustering of compounds in chemical space. A two-dimensional representation of chemical space being partitioned into clusters of similar compounds using a simple sphere exclusion method.

set of compounds. In some cases it is possible to consider the average similarity between compounds and optimize this as an objective function. However, this requires generation of an N by N similarity matrix, which may become prohibitively large as N increases.[93] Heuristic clustering methods are thus more commonly used.[93] Such methods include k-means clustering,[94] sphere exclusion,[95] directed sphere exclusion,[96] and maxmin.[97] The aim of such methods is that, for each selected molecule, no similar molecules are then selected. This is illustrated using a two-dimensional representation for a simple sphere exclusion method in Figure 3. The centroid molecules R, B, G, and Y represent all of the molecules within a similarity of greater than 0.2. Iterative selection in this chemical space will finally encompass all molecules. A secondary aim of compound selection is to pick clusters of two or more structurally similar compounds in each cluster, such that the initial assay results immediately provide some QSAR data to inform decision-making. In many cases the aim of compound selection is to augment an existing compound collection. In this case, the existing compound structures can be used as an input to the diverse selection algorithm. This can be used to select new compounds that "fill the gaps" in chemical space. Despite this usefulness of diversity selection methods, the use of virtual screening methods should always be considered in a resource constrained environment, with sufficient knowledge of the protein target and its structure. Both molecular docking[98] and pharmacophore analysis[99] can improve hit rates in HTS assays and are commonly used.

In summary, the process of selecting a representative subset of compounds from a large collection relies heavily on the ill-defined concept of molecular similarity. However, the concept is vital as it allows lead molecules to be identified at reduced cost and effort through hit identification and explosion.

■ CONCLUSIONS AND DISCUSSION

Shrewd selection of screening compounds is one of the most vital enabling steps in the drug development process. There are no strict rules, only rules of thumb. No compound filters are globally applicable, and no diversity metrics or selection methods can be proven as optimal. However, misapplication of filtering can reduce chemical diversity within a project and preclude many novel discoveries. Conversely, careful filtering reduces the risk of

false positives and downstream ADMET failures, while sensible compound selection can yield libraries that cover larger regions of chemical space and increase true positive hit rates. ADMET concerns may not be as important for chemical probes developed in academic groups, but solubility, cell permeability, and potential chemical reactivity are all still important considerations, and chemical diversity is still highly desirable. There are numerous sources of compound interference, which plague HTS assays. However, recent large-scale analyses have identified molecular scaffolds that appear as frequent hitters in numerous assays. The resultant data is very useful and should be incorporated either into library filtering or triaging of assay data. However, if every group used the same filters, then every group would test similar compounds and many useful molecules could be missed. Large screening libraries in industry include a substantial fraction of commercially available compounds. Thus, if an academic group sources from commercial vendors and uses traditional industry filters, then they will develop smaller relatives of the big industrial libraries with little or no chemical novelty. It may thus be advisable for academics to consider synthesizing or purchasing molecules in untapped regions of chemical space, particularly embodying multiple stereogenic centers, to maximize chemical diversity and increase the number of unique chemical entities tested. Diversity should also be maximized by considering natural products and biogenic scaffolds, which may show improved ADMET properties. At present, commercially available chemical space is heavily skewed toward flat compounds with many aromatic rings. While this makes synthesis more tractable, it excludes many sources of chemical diversity and shifts screening libraries away from biogenic scaffolds and toward pharmacological risks. These risks have been recently quantified and the results are compelling.[25] This problem will only be remedied by customers changing their practices to incentivise chemical suppliers.

A screening library must have the correct balance of molecular weight and log P, tailored to the constraints of the assay. Once a true positive hit has been identified, increasing size and complexity in tandem with lipophilicity is expected to increase both affinity and specificity. It is important to note that the ideal range of chemical and physicochemical properties of an HTS library differs when considering different assay platforms or protein targets. An optimal screening library for a fragment-based screen or targetting a protein−protein interaction will thus be different from a traditional kinase set and should be carefully designed. Due to the economies of scale with respect to purchasing a screening library, cost sharing between academic and government laboratories can increase the scope of screening efforts. Some companies may be willing to share portions of their screening libraries, in return for IP rights, on projects focused on commercially viable, validated targets. With respect to compound selection, there are numerous existing methods for measuring chemical similarity and selecting diverse sets of compounds, but no ideal metric can exist. While current work has highlighted the best applications of fingerprinting, shape-based, and pharmacophore methods, these are all evolving fields, and no technique can be proven superior in all cases. However, compound selection through analysis of molecular similarity reduces the size and cost of screening libraries while retaining diversity.

One question of great importance that has not been addressed in great detail is how many compounds need to be tested to ensure a sufficient coverage of chemical space.[100] This question can be answered by considering the number of lead series desired,

the false positive rate, the number of molecules assayed per cluster, and the hit rate of the primary screen. Such an analysis predicts that on average one lead series can be developed from testing approximately 350,000 diverse compounds in a typical HTS screen.[101] This number applies only to leads successfully developed into marketed drugs and is thus not appropriate when considering chemical probe discovery. However, it is commonly accepted that some targets are more druggable than others such that this value can vary greatly and that some screens will yield no successful lead series. Due to the importance of HTS in the development of new drugs and chemical probes, high-quality screening libraries are a key asset of any research group, and there are many factors to be weighed. However, each library will be unique and should be suited to the particular needs of the screening group. With the rapid increase in the number of purchasable molecules, the almost limitless volume of chemical space and the proliferation of HTS groups, rational selection of diverse hit-like compounds seems likely to continue as a lynchpin of drug development.

■ AUTHOR INFORMATION

Corresponding Author

*spring@ch.cam.ac.uk

■ ACKNOWLEDGMENT

■ KEYWORDS

Drug-like molecule: a molecule with molecular properties that overlap with the majority of existing drugs

Frequent hitter: a molecule or molecular substructure that hits numerous screening assays on different drug targets with a mode of action that is assumed to be non-specific

High-throughput screening: a screening process that utilises robotics and rapid data processing to perform millions of assays in a short space of time

Molecular diversity: a measure of how well a subset of molecules represents a larger set of molecules. A more diverse subset will tend to have a lower molecular similarity between molecules

Molecular similarity: a measure of the relatedness of two molecules. This would ideally quantify the similarity in biological effect but in practice tends to quantify the similarity in structure

Substructure filter: a computational filter used to remove molecules containing molecular substructures that are considered to give rise to non-specific binding or deleterious pharmacodynamic properties.

■ ABBREVIATIONS

ADMET: absorption, distribution, metabolism, elimination and toxicity

HTS: high-throughput screening
log P: octanol/water partition coefficient
PAINS: pan assay interference compounds
PSA: polar surface area
REOS: rapid elimination of swill

■ REFERENCES

(1) Pauling, L., Itano, H., Singer, S., and Wells, I. (1949) Sickle cell anemia, a molecular disease. *Science 110*, 543.

(2) Koehn, F. E., and Carter, G. T. (2005) The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discovery 4*, 206–220.

(3) Kennedy, J. P., Williams, L., Bridges, T. M., Daniels, R. N., Weaver, D., and Lindsley, C. W. (2008) Application of combinatorial chemistry science on modern drug discovery. *J. Comb. Chem. 10*, 345–354.

(4) Spring, D. R. (2005) Chemical genetics to chemical genomics: small molecules offer big insights. *Chem. Soc. Rev. 34*, 472–482.

(5) Kodadek, T. (2010) Rethinking screening. *Nat. Chem. Biol. 6*, 162–165.

(6) McCarthy, A. (2010) The NIH molecular libraries program: identifying chemical probes for new medicines. *Chem. Biol. 17*, 549–550.

(7) Workman, P., and Collins, I. (2010) Probing the probes: fitness factors for small molecule tools. *Chem. Biol. 17*, 561–577.

(8) Inglese, J., Johnson, R. L., Simeonov, A., Xia, M. H., Zheng, W., Austin, C. P., and Auld, D. S. (2007) High-throughput screening assays for the identification of chemical probes. *Nat. Chem. Biol. 3*, 466–479.

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

(10) Ajay, Walters, W. P., and Murcko, M. A. (1998) Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?. *J. Med. Chem. 41*, 3314–3324.

(11) Teague, S. J., Davis, A. M., Leeson, P. D., and Oprea, T. (1999) The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. 38*, 3743–3748.

(12) Lloyd, D. G., Golfis, G., Knox, A. J., Fayne, D., Meegan, M. J., and Oprea, T. I. (2006) Oncology exploration: charting cancer medicinal chemistry space. *Drug Discovery Today 11*, 149–159.

(13) Baurin, N., Baker, R., Richardson, C., Chen, I., Foloppe, N., Potter, A., Jordan, A., Roughley, S., Parratt, M., Greaney, P., Morley, D., and Hubbard, R. E. (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci. 44*, 643–651.

(14) Chuprina, A., Lukin, O., Demoiseaux, R., Buzko, A., and Shivanyuk, A. (2010) 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. 50*, 470–479.

(15) Editorial (2007) The academic pursuit of screening, *Nat. Chem. Biol. 3*, 433−433.

(16) Monge, A., Arrault, A., Marot, C., and Morin-Allory, L. (2006) Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol Divers 10*, 389–403.

(17) Sirois, S., Hatzakis, G., Wei, D. Q., Du, Q. S., and Chou, K. C. (2005) Assessment of chemical libraries for their druggability. *Computational Biology and Chemistry 29*, 55–67.

(18) Verheij, H. J. (2006) Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Diversity 10*, 377–388.

(19) Voigt, J. H., Bienfait, B., Wang, S. M., and Nicklaus, M. C. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci. 41*, 702–712.

(20) Irwin, J. J., and Shoichet, B. K. (2005) ZINC−a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model. 45*, 177–182.

(21) Bohacek, R. S., McMartin, C., and Guida, W. C. (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev. 16*, 3–50.

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

(23) Higueruelo, A. P., Schreyer, A., Bickerton, G. R. J., Pitt, W. R., Groom, C. R., and Blundell, T. L. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des. 74*, 457–467.

(24) Lovering, F., Bikker, J., and Humblet, C. (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem. 52*, 6752–6756.

(25) Ritchie, T. J., and Macdonald, S. J. (2009) The impact of aromatic ring count on compound developability—are too many aromatic rings a liability in drug design?. *Drug Discovery Today 14*, 1011–1020.

(26) Newman, D. J., and Cragg, G. M. (2007) Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod. 70*, 461–477.

(27) Singh, N., Guha, R., Giulianotti, M. A., Pinilla, C., Houghten, R. A., and Medina-Franco, J. L. (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model. 49*, 1010–1024.

(28) Feher, M., and Schmidt, J. M. (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci. 43*, 218–227.

(29) Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J., and Shoichet, B. K. (2009) Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol. 5*, 479–483.

(30) Spandl, R. J., Bender, A., and Spring, D. R. (2008) Diversity-oriented synthesis; a spectrum of approaches and results. *Org. Biomol. Chem. 6*, 1149–1158.

(31) Clemons, P. A., Bodycombe, N. E., Carrinski, H. A., Wilson, J. A., Shamji, A. F., Wagner, B. K., Koehler, A. N., and Schreiber, S. L. (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U.S.A. 107*, 18787–18792.

(32) Hann, M. M., Leach, A. R., and Harper, G. (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci. 41*, 856–864.

(33) Dobson, P. D., and Kell, D. B. (2008) Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule?. *Nat. Rev. Drug Discovery 7*, 205–220.

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

(35) Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem. 1*, 55–68.

(36) Veber, D. F. (2003) Molecular properties that influence the oral bioavailability of drug candidates. *Abstr. Papers Am. Chem. Soc. 225*, U208–U208.

(37) Walters, W. P., and Murcko, M. A. (2000) Library filtering systems and prediction of drug-like properties, in *Virtual Screening for Bioactive Molecules (Methods and Principles in Medicinal Chemistry)*, pp 15–30, Wiley-VCH, Weinheim.

(38) Oprea, T. I. (2000) Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des. 14*, 251–264.

(39) Egan, W. J., Merz, K. M., and Baldwin, J. J. (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem. 43*, 3867–3877.

(40) Lee, M. L., and Schneider, G. (2001) Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *J. Comb. Chem. 3*, 284–289.

(41) Martin, Y. C. (2005) A bioavailability score. *J. Med. Chem. 48*, 3164–3170.

(42) Palm, K., Stenberg, P., Luthman, K., and Artursson, P. (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res. 14*, 568–571.

(43) Subramanian, G., and Kitchen, D. B. (2006) Computational approaches for modeling human intestinal absorption and permeability. *J. Mol. Model. 12*, 577–589.

(44) Johnson, T. W., Dress, K. R., and Edwards, M. (2009) Using the Golden Triangle to optimize clearance and oral absorption. *Bioorg. Med. Chem. Lett. 19*, 5560–5564.

(45) Leeson, P. D., and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery 6*, 881–890.

(46) Hughes, J. D., Blagg, J., Price, D. A., Bailey, S., DeCrescenzo, G. A., Devraj, R. V., Ellsworth, E., Fobian, Y. M., Gibbs, M. E., Gilles, R. W., Greene, N., Huang, E., Krieger-Burke, T., Loesel, J., Wager, T., Whiteley, L., and Zhang, Y. (2008) Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett. 18*, 4872–4875.

(47) Cronin, D., and Mark, T. (2006) The role of hydrophobicity in toxicity prediction. *Curr. Comput.-Aided Drug Des. 2*, 405–413.

(48) Wenlock, M. C., Austin, R. P., Barton, P., Davis, A. M., and Leeson, P. D. (2003) A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem. 46*, 1250–1256.

(49) Delaney, J. S. (2005) Predicting aqueous solubility from structure. *Drug Discovery Today 10*, 289–295.

(50) Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998) Virtual screening—an overview. *Drug Discovery Today 3*, 160–178.

(51) Weininger, D. (1988) Smiles, a chemical language and information-system 0.1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci. 28*, 31–36.

(52) Rishton, G. M. (1997) Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today 2*, 382–384.

(53) Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res. 34*, D668–D672.

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

(55) Editorial. (2009) Screening we can believe, *Nat. Chem. Biol. 5*, 127-127.

(56) Snodin, D. J. (2010) Genotoxic impurities: from structural alerts to qualification. *Org. Process Res. Dev. 14*, 960–976.

(57) Hopkins, A. L., Groom, C. R., and Alex, A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today 9*, 430–431.

(58) Hajduk, P. J., and Greer, J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery 6*, 211–219.

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

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

(61) Seidler, J., McGovern, S. L., Doman, T. N., and Shoichet, B. K. (2003) Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem. 46*, 4477–4486.

(62) Feng, B. Y., Simeonov, A., Jadhav, A., Babaoglu, K., Inglese, J., Shoichet, B. K., and Austin, C. P. (2007) A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem. 50*, 2385–2390.

(63) Huth, J. R., Mendoza, R., Olejniczak, E. T., Johnson, R. W., Cothron, D. A., Liu, Y. Y., Lerner, C. G., Chen, J., and Hajduk, P. J. (2005) ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc. 127*, 217–224.

(64) Simeonov, A., Jadhav, A., Thomas, C. J., Wang, Y., Huang, R., Southall, N. T., Shinn, P., Smith, J., Austin, C. P., Auld, D. S., and Inglese,

J. (2008) Fluorescence spectroscopic profiling of compound libraries. *J. Med. Chem. 51*, 2363–2371.

(65) Pearce, B. C., Sofia, M. J., Good, A. C., Drexler, D. M., and Stock, D. A. (2006) An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model. 46*, 1060–1068.

(66) Jadhav, A., Ferreira, R. S., Klumpp, C., Mott, B. T., Austin, C. P., Inglese, J., Thomas, C. J., Maloney, D. J., Shoichet, B. K., and Simeonov, A. (2009) Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem. 53*, 37–51.

(67) Axerio-Cilies, P., Castaneda, I. P., Mirza, A., and Reynisson, J. (2009) Investigation of the incidence of "undesirable" molecular moieties for high-throughput screening compound libraries in marketed drug compounds. *Eur. J. Med. Chem. 44*, 1128–1134.

(68) 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., and Shoichet, B. K. (2008) Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *J. Med. Chem. 51*, 2502–2511.

(69) Park, C. M., Bruncko, M., Adickes, J., Bauch, J., Ding, H., Kunzer, A., Marsh, K. C., Nimmer, P., Shoemaker, A. R., Song, X., Tahir, S. K., Tse, C., Wang, X. L., Wendt, M. D., Yang, X. F., Zhang, H. C., Fesik, S. W., Rosenberg, S. H., and Elmore, S. W. (2008) Discovery of an orally bioavailable small molecule inhibitor of prosurvival B-cell lymphoma 2 proteins. *J. Med. Chem. 51*, 6902–6915.

(70) Zhang, Q., and Muegge, I. (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem. 49*, 1536–1548.

(71) Misra, R. N., Xiao, H. Y., Kim, K. S., Lu, S. F., Han, W. C., Barbosa, S. A., Hunt, J. T., Rawlins, D. B., Shan, W. F., Ahmed, S. Z., Qian, L. G., Chen, B. C., Zhao, R. L., Bednarz, M. S., Kellar, K. A., Mulheron, J. G., Batorsky, R., Roongta, U., Kamath, A., Marathe, P., Ranadive, S. A., Sack, J. S., Tokarski, J. S., Pavletich, N. P., Lee, F. Y. F., Webster, K. R., and Kimball, S. D. (2004) N-(Cycloalkylamino)acyl-2-aminothiazole inhibitors of cyclin-dependent kinase 2. N-[5-[[[5-(1,1-dimethylethyl)-2-oxazolyl]methyl]thio]-2-thiazolyl]-4-piperidinecarboxamide (BMS-387032), a highly efficacious and selective antitumor agent. *J. Med. Chem. 47*, 1719–1728.

(72) Pak, D. T., and Sheng, M. (2003) Targeted protein degradation and synapse remodeling by an inducible protein kinase. *Science 302*, 1368–1373.

(73) Syed, N., Smith, P., Sullivan, A., Spender, L. C., Dyer, M., Karran, L., O'Nions, J., Allday, M., Hoffmann, I., Crawford, D., Griffin, B., Farrell, P. J., and Crook, T. (2006) Transcriptional silencing of Polo-like kinase 2 (SNK/PLK2) is a frequent event in B-cell malignancies. *Blood 107*, 250–256.

(74) Inglis, K. J., Chereau, D., Brigham, E. F., Chiou, S. S., Schobel, S., Frigon, N. L., Yu, M., Caccavello, R. J., Nelson, S., Motter, R., Wright, S., Chian, D., Santiago, P., Soriano, F., Ramos, C., Powell, K., Goldstein, J. M., Babcock, M., Yednock, T., Bard, F., Basi, G. S., Sham, H., Chilcote, T. J., McConlogue, L., Griswold-Prenner, I., and Anderson, J. P. (2009) Polo-like kinase 2 (PLK2) phosphorylates alpha-synuclein at serine 129 in central nervous system. *J. Biol. Chem. 284*, 2598–2602.

(75) de Carcer, G., Perez de Castro, I., and Malumbres, M. (2007) Targeting cell cycle kinases for cancer therapy. *Curr. Med. Chem. 14*, 969–985.

(76) Stein, R. L. (2003) High-throughput screening in academia: the Harvard experience. *J. Biomol. Screening 8*, 615–619.

(77) Austin, C. P., Brady, L. S., Insel, T. R., and Collins, F. S. (2004) NIH Molecular Libraries Initiative. *Science 306*, 1138–1139.

(78) Keseru, G. M., and Makara, G. M. (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discovery 8*, 203–212.

(79) Matter, H. (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem. 40*, 1219–1229.

(80) Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002) Do structurally similar molecules have similar biological activity?. *J. Med. Chem. 45*, 4350–4358.

(81) Rishton, G. M. (2008) Molecular diversity in the context of leadlikeness: compound properties that enable effective biochemical screening. *Curr. Opin. Chem. Biol. 12*, 340–351.

(82) Bender, A., and Glen, R. C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem. 2*, 3204–3218.

(83) Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci. 38*, 983–996.

(84) Malumbres, M., and Barbacid, M. (2007) Cell cycle kinases in cancer. *Curr. Opin. Genet. Dev. 17*, 60–65.

(85) *Sybyl*, Tripos, St. Louis, MO.

(86) Hert, J., Willett, P., and Wilton, D. J. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci. 44*, 1177–1185.

(87) Naylor, E., Arredouani, A., Vasudevan, S. R., Lewis, A. M., Parkesh, R., Mizote, A., Rosen, D., Thomas, J. M., Izumi, M., Ganesan, A., Galione, A., and Churchill, G. C. (2009) Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol. 5*, 220–226.

(88) Duan, J. X., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010) Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Model. 29*, 157–170.

(89) Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C. K., Glick, M., and Davies, J. W. (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Modell. 49*, 108–119.

(90) Steffen, A., Kogej, T., Tyrchan, C., and Engkvist, O. (2009) Comparison of molecular fingerprint methods on the basis of biological profile data. *J. Chem. Inf. Model. 49*, 338–347.

(91) Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today 11*, 1046–1053.

(92) Hawkins, P. C. D., Skillman, A. G., and Nicholls, A. (2007) Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem. 50*, 74–82.

(93) Willett, P. (1999) Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol. 6*, 447–457.

(94) Higgs, R. E., Bemis, K. G., Watson, I. A., and Wikel, J. H. (1997) Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci. 37*, 861–870.

(95) Hudson, B. D., Hyde, R. M., Rahr, E., and Wood, J. (1996) Parameter based methods for compound selection from chemical databases. *Quant. Struct.-Act. Relat. 15*, 285–289.

(96) Gobbi, A., and Lee, M. L. (2003) DISE: Directed Sphere Exclusion. *J. Chem. Inf. Comput. Sci. 43*, 317–323.

(97) Schmuker, M., Givehchi, A., and Schneider, G. (2004) Impact of different software implementations on the performance of the Maxmin method for diverse subset selection. *Mol. Diversity 8*, 421–425.

(98) Shoichet, B. K., McGovern, S. L., Wei, B. Q., and Irwin, J. J. (2002) Lead discovery using molecular docking. *Curr. Opin. Chem. Biol. 6*, 439–446.

(99) Wolber, G., and Langer, T. (2005) LigandScout: 3-d pharmacophores derived from protein-bound Ligands and their use as virtual screening filters. *J. Chem. Inf. Model. 45*, 160–169.

(100) Lipkin, M. J., Stevens, A. P., Livingstone, D. J., and Harris, C. J. (2008) How large does a compound screening collection need to be?. *Comb. Chem. High Throughput Screening 11*, 482–493.

(101) Harper, G., Pickett, S. D., and Green, D. V. (2004) Design of a compound screening collection for use in high throughput screening. *Comb. Chem. High Throughput Screening 7*, 63–70.

217

dx.doi.org/10.1021/cb100420r |*ACS Chem. Biol.* 2011, 6, 208–217